Gaussian process regression models for the analysis of survival data with competing risks, interval censoring and high dimensionality

Barrett, James Edward

Awarding institution:
King's College London

Gaussian process regression models for the analysis of survival data with competing risks, interval censoring and high dimensionality

James Barrett

Department of Mathematics
King's College London

Thesis submitted to King's College London for
the degree of Doctor of Philosophy

Abstract

We develop novel statistical methods for analysing biomedical survival data based on Gaussian process (GP) regression. GP regression provides a powerful non-parametric probabilistic method of relating inputs to outputs. We apply this to survival data which consist of time-to-event and covariate measurements. In the context of GP regression the covariates are regarded as 'inputs' and the event times are the 'outputs'. This allows for highly flexible inference of non-linear relationships between covariates and event times.

Many existing methods for analysing survival data, such as the ubiquitous Cox proportional hazards model, focus primarily on the hazard rate which is typically assumed to take some parametric or semi-parametric form. Our proposed model belongs to the class of accelerated failure time models and as such our focus is on directly characterising the relationship between the covariates and event times without any explicit assumptions on what form the hazard rates take. This provides a more direct route to connecting the covariates to survival outcomes with minimal assumptions. An application of our model to experimental data illustrates its usefulness.

We then apply multiple output GP regression, which can handle multiple potentially correlated outputs for each input, to competing risks survival data where multiple event types can occur. In this case the multiple outputs correspond to the time-to-event for each risk. By tuning one of the model parameters we can control the extent to which the multiple outputs are dependent thus allowing the specification of correlated risks. However, the identifiability problem, which states that it is not possible to infer whether risks are truly independent or otherwise on the basis of observed data, still holds. In spite of this fundamental limitation simulation studies suggest that in some cases assuming dependence can lead to more accurate predictions.

The second part of this thesis is concerned with high dimensional survival data where there are a large number of covariates compared to relatively few individuals. This leads to the problem of overfitting, where spurious relationships are inferred from the data. One strategy to tackle this problem is dimensionality reduction. The Gaussian process latent variable model (GPLVM) is a powerful method of extracting a low dimensional representation of high dimensional data. We extend the GPLVM to incorporate survival outcomes by combining

the model with a Weibull proportional hazards model (WPHM). By reducing the ratio of covariates to samples we hope to diminish the effects of overfitting.

The combined GPLVM-WPHM model can also be used to combine several datasets by simultaneously expressing them in terms of the same low dimensional latent variables. We construct the Laplace approximation of the marginal likelihood and use this to determine the optimal number of latent variables, thereby allowing detection of intrinsic low dimensional structure. Results from both simulated and real data show a reduction in overfitting and an increase in predictive accuracy after dimensionality reduction.

Acknowledgements

I am deeply grateful to my supervisor Ton Coolen for all of his support, encouragement and patience (and his boundless supply of coffee). Under his guidance I have learned to enjoy difficult mathematical problems. I would like to acknowledge all of my colleagues at the Institute for Mathematical and Molecular Biomedicine who created an environment where I could develop as a researcher. I have benefited incalculably from their knowledge and experience. Numerous meetings and discussions with the Disordered Systems group at the Department of Mathematics have helped me to refine my thoughts which I am thankful for. I would like to thank Tony Ng for his faith in my abilities and for keeping me busy with interesting problems and Katherine Lawler for her encouragement. Especially when I was starting out. I am very grateful to my family and friends for all of their support and timely distractions. For everything else I want to thank my wife Sonam who's patience and belief sustained me through the difficult times and who's companionship made the good times so enjoyable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis is to develop new statistical and computational tools to analyse biomedical survival data. The quest to accurately characterise individual patients, to identify disease subtypes, to determine the most appropriate treatment and ultimately to predict survival outcomes has lead researchers to gather vast quantities of information from patients at a molecular level. Experimental advances allow unprecedented access to the innermost structure and dynamics of cells, with the behaviour of individual genes and proteins open to measurement. Huge quantities of data can be routinely generated that offer a rich source of information about the underlying biological processes. These massive datasets allow us to build up a picture of how a cell works and offer clues as to what goes wrong with diseases such as cancer. Despite the promise those new types of data hold, they also provide serious challenges in terms of statistical analysis and biological interpretation.

Firstly, the scale of the underlying biological systems is challenging. Tens of thousands of genes and proteins form a complex network of interactions evolving in time that is beyond the ability of the human mind to comprehend or understand without the aid of computational models or mathematical tools. There is therefore a need to develop statistical tools that can analyse these huge volumes of data and extract information that is of practical use and offers some genuine biological insight. The sheer number of measurements that can be acquired for one individual (currently up to hundreds of thousands) poses serious challenges to their statistical analysis due to the risk of overfitting where spurious relationships are inferred between data that exist due to chance and fail to occur in validation data. The greater the number of covariates compared to the number of samples then the greater the danger that overfitting presents. It is not uncommon to have datasets with thousands of covariates and

less than a hundred patients. It is difficult to infer statistical relationships between these data because the large number of covariates leads to a large number of possible patterns but we have too few examples to learn from. The challenge is to robustly extract statistical relationships between these data in a way that avoids inferring spurious patterns or structure.

A second problem posed by current biomedical data is that they come from highly non-linear systems. Genes and proteins form a complex network of interactions that are not always intuitively easy to understand or interpret. Whereas in the past clinical covariates such as age or sex may have had a monotonic relationship with survival outcomes this may no longer be the case with the types of covariates currently available. The dysregulation of particular genes or functional gene groups, or the interaction of certain proteins may be related to survival outcomes in a non-monotonic fashion. New mathematical tools need to be sufficiently flexible to accommodate potentially complicated non-linear relationships.

Finally, despite impressive experimental advances our ability to measure what is happening at any given time at a molecular level remains limited by a variety of factors. Data are often noisy and limited in spatial or temporal resolution. There may be heterogeneity within cohorts or within individuals that is not captured by the covariates. Datasets may suffer from missing values or be contaminated due to batch effects or other confounding factors. When survival data are acquired it is important to ask if competing risks are present and to consider whether censoring occurs independently of the risks under study. New statistical methods should be designed with these considerations in mind.

The aim of this thesis is to develop mathematical tools based on Gaussian process (GP) regression that attempt to address some of the issues raised above. In the first part of this thesis we develop GP regression models for survival data. The motivation behind this approach is to develop a flexible non-parametric probabilistic model that can handle non-monotonic survival data. We begin by applying this to the case of a single risk with independent censoring in Chapter 3.

Many existing methods of survival analysis focus on the hazard rate. Cox's proportional hazards model (Cox, 1972) is arguably the most popular such approach. These methods typically assume that the hazard rate splits into two components, one that captures the time effects and one that captures the covariate effects. Cox's model further assumes that the covariate effects are linear. It is not obvious however that this factorisation is always appropriate. In this work we will develop an accelerated failure time model where the event times are written as an unknown (and noise corrupted) function of the covariates. GP regression will be used to infer the unknown function in a flexible and non-parametric manner. From

this point of view the event times are considered 'outputs' and the covariates 'inputs' in a regression model. We argue that this approach is a more direct way of connecting the quantities that we have experimental access to, namely the covariates and the event times. Hazard rate models on the other hand take a more indirect route and need to somehow capture both the time and covariate effects on survival outcomes whereas our approach need only capture the covariate effects and consequently fewer assumptions are required.

Our model can also incorporate any type of censored and truncated observations relatively easily. In addition, we obtain estimates of when the event would have occurred to individuals that were censored. We perform several simulation studies which illustrate the models ability to infer non-monotonic relationships between the covariates and event times. We compare our model to more traditional models such as the Cox proportional hazards model, the Weibull proportional hazards model and a third model that is also based on GP regression but assumes a hazard rate similar to the Cox model but with non-linear covariate effects. We also apply our approach to some real data.

Throughout this work we will infer all model parameters in a Bayesian manner. We compute the maximum a posteriori (MAP) solution by numerically maximising the posterior density over parameters. We then construct the Laplace approximation of the hyperparameter posterior and from that numerically compute the MAP solution for hyperparameters. Some background theory on Bayesian inference is given in Chapter 2.

This work is extended to the competing risks scenario using multiple output GP regression in Chapter 4. Multiple output GP regression was originally developed for situations where multiple outputs are available corresponding to given inputs where the outputs may be statistically dependent. Again, we regard the time-to-event for different risks as the 'multiple outputs' and the covariates as the 'inputs'. Usually multiple output GP regression would be applied to data where all of the outputs corresponding to each input may be observed. There are two features of competing risks data that are interesting in this context. Firstly, at most one 'output' is available for each individual and secondly, once one of the outputs is observed we know that the remaining outputs must be greater than the observed output. This is because we know that remaining events would have occurred after the reported event time.

The model can assume either independent risks or dependent risks by tuning the value of one parameter. Of course, the identifiability problem (Tsiatis, 1975) means we cannot conclude whether the risks are truly independent or not in reality. Nevertheless, within the framework of the model we will infer the value of the parameter that best explains the observed

data. If the assumption of dependence has a higher probability then the model will follow this and exploit it to potentially make more accurate predictions. Consider, for example, two strongly dependent risks. If there is a region of the covariate space where only the first event has been observed we can still make accurate predictions of when the second type of event would occur for new individuals. This is because we know the second risk will behave similarly to the first risk. We also examine the issue of what happens in the hypothetical scenario where we 'disable' or 'switch off' one or more risks.

The second part of this thesis focuses on the problem of high dimensional data. The so-called 'curse of dimensionality' refers to the challenge of extracting genuine statistical patterns from such large volumes of data. This typically leads to the phenomenon of overfitting where statistical models tend to fit training data very well but fail to generalise to unseen data (sometimes called validation or test data). One strategy for dealing with this problem is dimensionality reduction which attempts to represent the information in terms of a smaller number of variables, often referred to as *latent variables*. One popular method of doing this is the Gaussian process latent variable model (GPLVM) developed by Lawrence (2005). The model attempts to generate a non-linear low dimensional representation of high dimensional data in terms the latent variables and since it is based on GP regression has some attractive features. It is non-parametric, probabilistic and can specify non-linear relationships between the low and high dimensional spaces. In Chapter 5 we overcome some mathematical challenges to construct the Laplace approximation of the marginal likelihood and use this for the purposes of dimensionality detection, by which we mean determining the most probable number of latent variables. We discuss results from numerical simulations that illustrate the effects of overfitting and the ability of the model to mitigate these.

In Chapter 6 we extend the GPLVM to incorporate high dimensional survival data. Our aim is that by reducing the dimension we can diminish the effects of overfitting and provide a more robust method of analysing survival data. We achieve this by coupling the GPLVM to a Weibull proportional hazards model. The low dimensional representation that is extracted will hopefully contain information that is relevant to the survival outcomes (in contrast to the unsupervised GPLVM which does not take survival outcomes into account). By constraining ourselves to a low dimensional space we aim to diminish the effects of overfitting and increase predictive accuracy. We study the effect of overfitting by generating simulated high dimensional data and test the ability of the model to reduce these effects. Finally, we apply the model to some experimental gene expression data and discuss these results.

Chapter 2

Background

In this chapter we provide the basic background theory and definitions for survival analysis, Bayesian inference, and Gaussian process regression and establish some notation. Several of these formulae will be used throughout this thesis so it will be useful to collect them here for reference.

2.1 Survival analysis

We give an overview of all the required survival analysis formulae and in particular look at the case of independent risks and the case of a single risk with independent censoring since both of these will be relevant later. We take as our starting point the joint event time density and derive other quantities such as hazard rates and the survival function from that.

2.1.1 Basic definitions

We will be interested in data of the form $\{(\mathbf{x}_1, \tau_1, \Delta_1), \ldots, (\mathbf{x}_N, \tau_N, \Delta_N)\}$ where $i = 1, \ldots, N$ with N individuals in total, and $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of covariates from individual i. The time until the first event is $\tau_i \geq 0$. We also require an indicator variable $\Delta_i = \{0, \ldots, R\}$ which tells us which of the R possible events occurred first. We reserve $\Delta_i = 0$ to denote censoring. Usually, one of the risks is called the primary risk and following convention we use $\Delta_i = 1$ to label it.

We will assume throughout that $\mathbf{x}_i$ does not depend on time. We also assume that only one event can occur and that the first event precludes the observation of any subsequent events that may have happened. Unless otherwise stated censoring is assumed to be random, that is,

the time of censoring is statistically independent from the other event times[1]. For example, we may study a cohort of cancer patients where the primary event is metastasis. We record for each patient the time until metastasis. A patient who dies from an unrelated disease, or doesn't return to the clinic for some reason (such as emigration) would be considered right censored.

There are several other types of censoring. An individual is *left censored* if the event is known to have occurred before a certain time. *Interval censoring* means that the event is known to occur within a certain interval of time. We will later apply our model to interval censored data. Instead of a single event time we would now have a pair (τ_i^l, τ_i^u) which define the lower and upper bounds of the censoring interval respectively.

An individual who is only included in a study if the event occurred before a known time is considered *right truncated*. If an an individual's inclusion in a study is conditional on the event occurring after a certain time then that individual is said to be *left truncated*. *Interval truncation* refers to the case where only individuals who experience the event within a certain interval are included. Each type of censoring corresponds to a different contribution to the data likelihood which we will give below.

We assume there exist R random variables corresponding to the event times for each event. The statistical properties of these event times are completely specified by the *joint event time density* which is denoted by

$$p_i(\tau_0, \ldots, \tau_R) \quad \text{for } i = 1, \ldots, N. \tag{2.1}$$

For the sake of generality we include censoring since it can be regarded as another risk. The integrated event time density is

$$S_i(\tau_0, \ldots, \tau_R) = \int_{\tau_0}^{\infty} \mathrm{d}s_0 \cdots \int_{\tau_R}^{\infty} \mathrm{d}s_R \, p_i(s_0, \ldots, s_R). \tag{2.2}$$

The *survival function* gives the probability that individual i will be alive at time τ

$$S_i(\tau) = S_i(\tau_0, \ldots, \tau_R)\Big|_{\tau_q = \tau \text{ for } q = 0, \ldots, R}. \tag{2.3}$$

[1] If censoring is not independent then it can be treated as a competing risk and handled within the formalism for competing risks developed in Chapter 4.

The *cause-specific hazard rate* for risk r is

$$\pi_i^r(\tau) = \frac{\left(\prod_{q \neq r} \int_\tau^\infty \mathrm{d}s_q\right) p_i(s_0, \ldots, s_{r-1}, \tau, s_{r+1}, \ldots, s_R)}{S_i(\tau)} \tag{2.4}$$

and gives the probability that event r will occur in the interval $[\tau, \tau + \mathrm{d}\tau)$ given that no event has occurred up until time τ. If we multiply both sides of (2.4) by $S_i(\tau)$ then the right hand side gives us the probability that event type $\Delta_i = r$ will be reported at time τ for individual i. We call this the *individual data likelihood* and denote it as

$$P_i(\tau, r) = \pi_i^r(\tau) S_i(\tau). \tag{2.5}$$

It is possible to write the survival function in terms of the hazard rates. We first note that (2.4) can be written as

$$\pi_i^r(\tau) = -\frac{\partial}{\partial \tau_r} \log S_i(\tau_0, \ldots, \tau_R)\Big|_{\tau_q = \tau \text{ for } q = 0, \ldots, R}. \tag{2.6}$$

We can then use this as follows:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \log S_i(\tau) &= \sum_{q=0}^{R} \frac{\partial}{\partial \tau_q} \log S_i(\tau_0, \ldots, \tau_R)\Big|_{\tau_q = \tau \text{ for } q = 0, \ldots, R} \\ &= -\sum_{q=0}^{R} \pi_i^q(\tau). \end{aligned} \tag{2.7}$$

From this we obtain

$$S_i(\tau) = e^{-\sum_{q=0}^{R} \int_0^\tau \pi_i^q(s) \mathrm{d}s}. \tag{2.8}$$

Another quantity that is sometimes used is the *cumulative incidence function* which gives the probability that $\Delta_i = r$ for individual i and that event r will occur before time τ:

$$\begin{aligned} C_i^r(\tau) &= \int_0^\tau \mathrm{d}s \, P_i(s, r) \\ &= \int_0^\tau \mathrm{d}s \, \pi_i^r(s) e^{-\sum_{q=0}^{R} \int_0^s \mathrm{d}s' \pi_i^q(s')}. \end{aligned} \tag{2.9}$$

Note that both the survival function (2.8) and cumulative incidence function depend (2.9) on all of the hazard rates.

2.1.2 Data likelihood

We now provide a derivation of the individual data likelihood in the case of independent right censoring. Let $\tau_i^0, \ldots, \tau_i^R$ be the times at which each event occurs in the hypothetical scenario where each event can be observed without precluding the observation of any subsequent events. To calculate $P_i(\tau_i, \Delta_i)$ we note the following:

$$(\tau_i, \Delta_i) = \begin{cases} (\tau_i^0, 0) & \text{if } \tau_i^0 < \tau_i^1, \ldots, \tau_i^0 < \tau_i^R \\ \vdots & \vdots \\ (\tau_i^r, r) & \text{if } \tau_i^r < \tau_i^0, \ldots, \tau_i^r < \tau_i^{r-1}, \tau_i^r < \tau_i^{r+1}, \ldots, \tau_i^r < \tau_i^R \\ \vdots & \vdots \\ (\tau_i^R, R) & \text{if } \tau_i^R < \tau_i^0, \ldots, \tau_i^R < \tau_i^{R-1}. \end{cases}$$

In what follows we will use the *Kronecker delta function* which is defined by

$$\delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \tag{2.10}$$

and the *Step function* (or *Heaviside function*)

$$\theta(x - y) = \begin{cases} 1 & \text{for } x > y \\ 0 & \text{for } x < y \\ 1/2 & \text{for } x = y. \end{cases} \tag{2.11}$$

We will also use the *Dirac delta function* which can roughly be described by

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.12}$$

with the following properties:

$$\int \mathrm{d}x\, \delta(x) = 1 \tag{2.13}$$

$$\int \mathrm{d}x\, f(x)\delta(x) = f(0). \tag{2.14}$$

We now integrate over the joint event time distribution to get

$$\begin{aligned} P_i(\tau_i, \Delta_i) &= \int_0^\infty \cdots \int_0^\infty \mathrm{d}s_i^0 \cdots \mathrm{d}s_i^R \, p(s_i^0, \ldots, s_i^R) \left(\sum_{r=0}^R \delta_{\Delta_i, r} \delta(\tau_i - s_i^r) \prod_{q \neq r} \theta(s_i^q - s_i^r) \right) \\ &= \sum_{r=0}^R \delta_{\Delta_i, r} \int_0^\infty \mathrm{d}s_i^r \left(\prod_{q \neq r} \int_{s_i^r}^\infty \mathrm{d}s_i^q \right) p(s_i^0, \ldots, s_i^R) \delta(\tau_i - s_i^r) \\ &= \sum_{r=0}^R \delta_{\Delta_i, r} \left(\prod_{q \neq r} \int_{\tau_i}^\infty \mathrm{d}s_i^q \right) p(s_i^0, \ldots, s_i^{r-1}, \tau_i, s_i^{r+1}, \ldots, s_i^R). \end{aligned} \tag{2.15}$$

We can use (2.4, 2.8) to rewrite this expression in terms of hazard rates:

$$\begin{aligned} P_i(\tau_i, \Delta_i) &= \sum_{r=0}^R \delta_{\Delta_i, r} \pi_i^r(\tau_i) S_i(\tau_i) \\ &= \sum_{r=0}^R \delta_{\Delta_i, r} \pi_i^r(\tau_i) e^{-\sum_{q=0}^R \int_0^{\tau_i} \mathrm{d}s \pi_i^q(s)}. \end{aligned} \tag{2.16}$$

The overall data likelihood factorises over samples:

$$p(D | \{\pi_i^1, \ldots, \pi_i^R\}_{i=1,\ldots,N}) = \prod_{i=1}^N P_i(\tau_i, \Delta_i). \tag{2.17}$$

We can view the likelihood in a slightly different way by writing it as

$$\begin{aligned} p(D | \{\pi_i^1, \ldots, \pi_i^R\}_{i=1,\ldots,N}) &= \prod_{i=1}^N \pi_i^{\Delta_i}(\tau_i) S_i(\tau_i) \\ &= \prod_{r=0}^R \left\{ \prod_{i=1}^N [\pi_i^r(\tau_i)]^{\delta_{\Delta_i, r}} e^{-\int_0^{\tau_i} \mathrm{d}s \, \pi_i^r(s)} \right\}. \end{aligned} \tag{2.18}$$

This has the following interpretation. For risk r each individual will make one of two possible contributions. If event r occurred to an individual then they will contribute with $\pi_i^r(\tau_i) S_i(\tau_i)$. Otherwise they contribute with $S_i(\tau_i)$ which is the probability that no event has occurred up to that time. This is equivalent to regarding all individuals who didn't experience risk r as right censored.

2.1.3 Independent risks and identifiability

Risks are independent if the joint event time density factorises

$$p_i(\tau_0, \ldots, \tau_R) = p_i^0(\tau_0) \cdots p_i^R(\tau_R). \tag{2.19}$$

Consequently the survival function factorises over risks

$$S_i(\tau) = S_i^0(\tau) \cdots S_i^R(\tau) \tag{2.20}$$

where

$$S_i^r(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i^r(s). \tag{2.21}$$

The cause specific hazard rates simplify to

$$\pi_i^r(\tau) = p_i^r(\tau)/S_i^r(\tau). \tag{2.22}$$

The survival function can now be written in terms of the hazard function as

$$S_i^r(\tau) = e^{-\int_0^\tau \mathrm{d}s \pi_i^r(s)}. \tag{2.23}$$

If we equate this expression for $S_i^r(\tau)$ with (2.21) we can derive an expression for the marginal event time density:

$$\begin{aligned} \int_\tau^\infty \mathrm{d}s\, p_i^r(s) &= e^{-\int_0^\tau \mathrm{d}s \pi_i^r(s)} \\ \frac{\mathrm{d}}{\mathrm{d}\tau}\left[1 - \int_0^\tau \mathrm{d}s\, p_i^r(s)\right] &= \frac{\mathrm{d}}{\mathrm{d}\tau} e^{-\int_0^\tau \mathrm{d}s \pi_i^r(s)} \\ p_i^r(\tau) &= \pi_i^r(\tau) e^{-\int_0^\tau \mathrm{d}s\, \pi_i^r(s)}. \end{aligned} \tag{2.24}$$

The data likelihood also simplifies and we can write (2.18) as

$$p(D|\pi_i^1, \ldots, \pi_i^R) = \prod_{r=0}^{R} \left\{ \prod_{i=1}^{N} [\pi_i^r(\tau_i)]^{\delta_{\Delta_i, r}} S_i^r(\tau_i) \right\}. \tag{2.25}$$

The identifiability problem (Tsiatis, 1975) states that is not possible to conclude on the basis of observed survival data alone whether or not the risks are independent. An easy way

to see this is to consider risks that are in fact dependent and described by a joint event time density $p_i(\tau_0, \ldots, \tau_R)$ that doesn't factorise. The hazard rates can then be obtained via (2.4). Given the hazard rates it is now possible to construct an alternative event time density with independent risks:

$$p_i(\tau_0, \ldots, \tau_R) = \prod_{r=0}^{R} \left[\pi_i^r(\tau) e^{-\int_0^\tau \mathrm{d}s\, \pi_i^r(s)} \right]. \tag{2.26}$$

According to (2.22) the hazard rates corresponding to (2.26) are exactly the same but the underlying event time density is very different. It is therefore always possible to find at least one alternative explanation of dependent risks that will be equally consistent with the observed data.

The only quantities which can be inferred from survival data are the hazard rates and any quantities which can be written in terms of the hazard rates such as the survival function (2.8) and cumulative incidence function (2.9). Quantities such as the joint event time density are unobservable and it is therefore impossible to infer whether the risks are statistically independent. Of course, it is possible that one explanation is more *plausible* than another but this must be based on external information such as prior knowledge of the system or intuition.

2.1.4 A single risk with independent censoring

In the specific case of a a single risk with independent censoring the above formulae simplify. The hazard rate is given by

$$\pi_i(\tau) = \frac{p_i(\tau)}{S_i(\tau)} \tag{2.27}$$

with

$$S_i(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i(s). \tag{2.28}$$

The individual data likelihood under right censoring is

$$P_i(\tau_i, \Delta_i) = \begin{cases} \pi_i(\tau) e^{-\int_0^\tau \mathrm{d}s\, \pi_i(s)} & \text{if } \Delta_i = 1 \\ e^{-\int_0^\tau \mathrm{d}s\, \pi_i(s)} & \text{if } \Delta_i = 0. \end{cases} \tag{2.29}$$

We can also include other types of censoring. The likelihood contributions corresponding to different censoring and truncation regimes are (see Klein and Moeschberger (2003, Section

3.5)):

$$\begin{aligned}
\text{primary event time known:} &\quad p(t_i) \\
\text{right censored:} &\quad S(t_i) \\
\text{left censored:} &\quad 1 - S(t_i) \\
\text{interval censored:} &\quad S(t_i^l) - S(t_i^u) \\
\text{left truncated:} &\quad p(t_i)/S(t_i^l) \\
\text{right truncated:} &\quad p(t_i)/(1 - S(t_i^u)) \\
\text{interval truncated:} &\quad p(t_i)/(S(t_i^u) - S(t_i^l)).
\end{aligned} \tag{2.30}$$

The cumulative incidence function (2.9) becomes

$$\begin{aligned}
C_i(\tau, 1) &= \int_0^\tau \mathrm{d}s\, \pi_i(s) e^{-\int_0^s \mathrm{d}s'\, \pi_i(s')} \\
&= -e^{-\int_0^s \mathrm{d}s'\, \pi_i(s')}\Big|_0^\tau \\
&= 1 - S_i(\tau).
\end{aligned} \tag{2.31}$$

From (2.24) we can also write the event time density in terms of the hazard rate

$$p_i(\tau) = \pi_i(\tau) e^{-\int_0^\tau \mathrm{d}s\, \pi_i(s)}. \tag{2.32}$$

Since we will develop a model for interval censored data we will provide a derivation of the corresponding individual likelihood term in (2.30). Again, we let τ_i^0 and τ_i^1 denote the time until right censoring and the time until the primary event that would be observed in the hypothetical world where both events can be observed. If an individual is interval censored we will observe a pair of times that define the interval $(\tau_i^l, \tau_i^u, \Delta_i = 1)$. This will be reported if $\tau_i^l \leq \tau_i^1 < \tau_i^u$ and $\tau_i^1 < \tau_i^0$ and conversely $(\tau_i^0, \Delta_i = 0)$ will be reported if $\tau_i^0 < \tau_i^1$. We then

integrate over the event time density:

$$\begin{aligned}
p(\tau_i, \Delta_i) &= \int_0^\infty \mathrm{d}s_i^0 \int_0^\infty \mathrm{d}s_i^1 \, p(s_i^0) p(s_i^1|f_i) \left\{ \delta_{\Delta_i,0} \delta(s_i^0 - \tau_i^0) \theta(s_i^1 - s_i^0) \right. \\
&\qquad \left. + \delta_{\Delta_i,1} \theta(s_i^1 - \tau_i^l) \theta(\tau_i^u - s_i^1) \theta(s_i^0 - s_i^1) \right\} \\
&= \delta_{\Delta_i,0} \int_0^\infty \mathrm{d}s_i^0 \int_{s_i^0}^\infty \mathrm{d}s_i^1 \, p(s_i^0) p(s_i^1|f_i) \delta(s_i^0 - \tau_i^0) + \delta_{\Delta_i,1} \int_{s_i^1}^\infty \mathrm{d}s_i^0 \, p(s_i^0) \int_{\tau_i^l}^{\tau_i^u} \mathrm{d}s_i^1 \, p(s_i^1|f_i) \\
&\propto \delta_{\Delta_i,0} \int_{\tau_i^0}^\infty \mathrm{d}s_i^1 \, p(s_i^1|f_i) + \delta_{\Delta_i,1} \int_{\tau_i^l}^{\tau_i^u} \mathrm{d}s_i^1 \, p(s_i^1|f_i) \\
&= \delta_{\Delta_i,0} S(\tau_i^0|f_i) + \delta_{\Delta_i,1} [S(\tau_i^u|f_i) - S(\tau_i^l|f_i)]. \qquad (2.33)
\end{aligned}$$

We have dropped terms that are independent of f_i.

2.2 Bayesian inference

Having observed data D we may wish to infer the values of various parameters or perhaps determine what the most appropriate choice of model is to explain the data. In a Bayesian approach we distinguish three levels of quantities that we can infer from the observed data:

- Microscopic data generating parameters $\mathbf{w}$ that typically scale with the number of samples in a dataset or the dimension of the dataset. In a linear regression model, for example, these would be the regression coefficients.
- Hyperparameters $\boldsymbol{\theta}_1$ that control qualitative features of the model such as the overall noise level. This level also includes hyperparameters $\boldsymbol{\theta}_2$ that control the distribution of microscopic parameters.
- Models H. For example, we may want to know what the most appropriate choice of model from a certain class is (such as what kernel function is the best choice in a Gaussian process).

For each level of uncertainty we obtain posterior densities using Bayes' Theorem:

$$p(\mathbf{w}|D, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H) = \frac{p(D|\mathbf{w}, \boldsymbol{\theta}_1, H)p(\mathbf{w}|\boldsymbol{\theta}_2, H)}{\int \mathrm{d}\mathbf{w}'\, p(D|\mathbf{w}', \boldsymbol{\theta}_1, H)\, p(\mathbf{w}'|\boldsymbol{\theta}_2, H)} \tag{2.34}$$

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|D, H) = \frac{p(D|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|H)}{\int \mathrm{d}\boldsymbol{\theta}'_1 \mathrm{d}\boldsymbol{\theta}'_2\, p(D|\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, H)p(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2|H)} \tag{2.35}$$

$$p(H|D) = \frac{p(D|H)p(H)}{\sum_{H'} p(D|H')p(H')}. \tag{2.36}$$

The likelihood terms are

$$p(D|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H) = \int \mathrm{d}\mathbf{w}\, p(D|\mathbf{w}, \boldsymbol{\theta}_1, H)\, p(\mathbf{w}|\boldsymbol{\theta}_2, H) \tag{2.37}$$

$$p(D|H) = \int \mathrm{d}\boldsymbol{\theta}_1 \mathrm{d}\boldsymbol{\theta}_2\, p(D|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|H). \tag{2.38}$$

Prior densities $p(\mathbf{w}|\boldsymbol{\theta}_1, H)$, $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|H)$ and $p(H)$ encode any prior knowledge or beliefs we may have about the parameters or models. If our beliefs are very weak then a broad prior should be chosen. The prior can also play a regularising role by preventing the inference of implausible parameters values (we will see examples of this in Chapter 6). Usually the prior over models is flat (that is, $p(H) = \textit{constant}$) unless we have a specific reason to favour one model over another before data are observed.

The probability density in (2.37) is called the *marginal density*. This integral is often analytically intractable and needs to be approximated. One such approximation is the Laplace approximation which we discus below.

The *maximum a posteriori* (MAP) solution is obtained by maximising the posterior density with respect to its argument. The MAP solution for the microscopic parameters for instance is

$$\hat{\mathbf{w}} = \mathrm{argmax}_{\mathbf{w}} p(\mathbf{w}|D, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H) \tag{2.39}$$

and provides a useful point estimate of the inferred parameters. Compare this to the *maximum likelihood* estimate of $\mathbf{w}$ which is obtained by solving $\max_{\mathbf{w}} p(D|\mathbf{w}, \boldsymbol{\theta}_1, H)$. The difference being that the prior term is included in the MAP estimate.

2.2.1 The Laplace approximation

As mentioned above the integral in (2.37) is often analytically intractable, and so we will construct a Gaussian approximation of $p(\mathbf{w}|D, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H)$. We drop H from our notation for

simplicity. We first define the negative log posterior likelihood

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\log p(\mathbf{w}|D, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H). \tag{2.40}$$

This is expanded to second order around $\hat{\mathbf{w}} = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$ using a Taylor expansion:

$$\mathcal{L}(\mathbf{w}) \approx \mathcal{L}(\hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}}) \cdot \mathbf{A}(\mathbf{w} - \hat{\mathbf{w}}) \tag{2.41}$$

where

$$\mathbf{A}_{ij} = \frac{\partial^2}{\partial w_i \partial w_j}\mathcal{L}(\mathbf{w})\Big|_{\hat{\mathbf{w}}}. \tag{2.42}$$

We can now rewrite (2.37) as

$$\begin{aligned} p(D|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \int \mathrm{d}\mathbf{w}\, e^{-N\mathcal{L}(\mathbf{w})} \\ &\approx \int \mathrm{d}\mathbf{w}\, e^{-N\mathcal{L}(\hat{\mathbf{w}}) - \frac{N}{2}(\mathbf{w}-\hat{\mathbf{w}})\cdot\mathbf{A}(\mathbf{w}-\hat{\mathbf{w}})} \\ &= p(D|\hat{\mathbf{w}}, \boldsymbol{\theta}_1)p(\hat{\mathbf{w}}|\boldsymbol{\theta}_2) \int \mathrm{d}\mathbf{w}\, e^{-\frac{N}{2}(\mathbf{w}-\hat{\mathbf{w}})\cdot\mathbf{A}(\mathbf{w}-\hat{\mathbf{w}})} \\ &= p(D|\hat{\mathbf{w}}, \boldsymbol{\theta}_1)p(\hat{\mathbf{w}}|\boldsymbol{\theta}_2)(2\pi)^{N/2}|(N\mathbf{A})^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|^{1/2}. \end{aligned} \tag{2.43}$$

The determinant of matrix $\mathbf{A}$ is denoted by $|\mathbf{A}|$. We can then define the approximated negative log hyperparameter likelihood as

$$\begin{aligned} \mathcal{L}_{hyp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= -\frac{1}{N}\log p(\mathbf{w}|D, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, H) \\ &= \mathcal{L}(\hat{\mathbf{w}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|N\mathbf{A}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|. \end{aligned} \tag{2.44}$$

2.3 Gaussian process regression

Gaussian process (GP) regression is a popular regression method that can be applied to standard regression and classification problems. It is a highly flexible non-parametric way of inferring an unknown function between sets of inputs and outputs. By specifying different *kernel functions* (defined below) we can infer a wide range of qualitatively different functions. The appeal of GP regression models lies in their conceptual simplicity and elegance. We outline the relevant definitions below. A superb introduction can be found in Rasmussen and

Williams (2006, Chapter 2).

2.3.1 Basic definitions

In a regression setting we have output variables $t_i \in \mathbb{R}$ associated with each of the observed covariate vectors $\mathbf{x}_i \in \mathbb{R}^d$ where $i = 1, \ldots, N$. The outputs are assumed to be given by some function of the inputs plus noise:

$$t_i = f(\mathbf{x}_i) + \xi_i \quad \text{for } i = 1, \ldots, N. \tag{2.45}$$

The noise is assumed to be Gaussian so $p(\xi_i) = \mathcal{N}(0, \beta^2)$. In Gaussian Process (GP) regression any finite collection of function values will be Gaussian distributed. A Gaussian process is defined by its mean and covariance functions:

$$m(\mathbf{x}) = \langle f(\mathbf{x}) \rangle \tag{2.46}$$

$$k(\mathbf{x}, \mathbf{x}') = \left\langle (f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}') \right\rangle. \tag{2.47}$$

This can be interpreted as a prior over functions. By generating any finite collection of test inputs we can can then generate a Gaussian random vector from the prior distribution whose elements are the function values at the test input locations. We typically write the collection of function values corresponding to the observed inputs as $\mathbf{f} = (f_1, \ldots, f_N)$ where $f_i = f(\mathbf{x}_i)$. In this case we can write

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}\mathbf{f}\cdot\mathbf{K}^{-1}\mathbf{f}}}{(2\pi)^{N/2}|\mathbf{K}|^{1/2}}. \tag{2.48}$$

The vector $\boldsymbol{\theta}$ contains hyperparameters required by the kernel function and the matrix $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) + \beta^2$. The kernel functions considered in this work are

$$\begin{aligned} &k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \mathbf{x}_i \cdot \mathbf{x}_j && \text{linear,} \\ &k(\mathbf{x}_i, \mathbf{x}_j) = \sigma(1 + \mathbf{x}_i \cdot \mathbf{x}_j)^2 && \text{polynomial (of second order),} \\ &k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp(-(\mathbf{x}_i - \mathbf{x}_j) \cdot \mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)/2) && \text{squared exponential.} \end{aligned} \tag{2.49}$$

In all three kernels the hyperparameter σ controls the variance of high dimensional outputs. The matrix $\mathbf{L} = \text{diag}(\mathbf{l})$. The components of $\mathbf{l} = (l_1^{-2}, \ldots, l_d^{-2})$ are known as *automatic relevance determination* (ARD) parameters and roughly tell us how important each covariate is. This is because l_μ defines a characteristic length scale over which the output associated with covariate μ varies. If the output varies a lot with a particular covariate then it is 'important'.

These hyperparameters are analogous to the coefficients in a linear regression model or the regression coefficients in Cox regression.

2.3.2 Inference and predictions

Having observed a dataset $D = \{(t_1, \mathbf{x}_1), \ldots, (t_N, \mathbf{x}_N)\}$ we may wish to make a prediction of the noise-free output f^* associated with a test input $\mathbf{x}^*$. The joint distribution of $\mathbf{f}$ and the noise-free test output f^* is

$$\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{k}^{*\mathrm{T}} \\ \mathbf{k}^* & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right) \tag{2.50}$$

where $\mathbf{k}^* \in \mathbb{R}^N$ and is defined by $\mathbf{k}^*_i = k(\mathbf{x}^*, \mathbf{x}_i)$ and $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. From this it can be shown that the conditional probability of f^* given the observed data and the test input is $p(f^*|D, \boldsymbol{\theta}) = \mathcal{N}(\mu, \kappa)$ with mean and variance:

$$\mu = \mathbf{k}^* \cdot \mathbf{K}^{-1}\mathbf{x} \tag{2.51}$$

$$\kappa = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^* \cdot \mathbf{K}^{-1}\mathbf{k}^*. \tag{2.52}$$

To obtain noisy predictions we simply change the variance to $\kappa + \beta^2$. The main task for 'fitting' or 'training' a GP regression model is to optimise the hyperparameters. This is done by maximising the marginal likelihood $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{K}(\boldsymbol{\theta}))$ with respect to the hyperparameters. For the sake of generality we have included a mean function $\boldsymbol{\eta}(\boldsymbol{\theta})$ which also depends on the hyperparameters.

Chapter 3

Gaussian process regression with one risk and independent censoring

3.1 Introduction

In this chapter we apply Gaussian process regression to survival data. We will confine ourselves to the case of a single risk with independent censoring. In the following chapter this will be extended to the competing risks problem. GP regression for survival analysis belongs to the class of accelerated failure time models. The event times are transformed so as to take negative and positive values and then the transformed event times are written as some function of the covariates plus a noise variable. In our case a GP prior is assumed for this function. This allows for a flexible non-parametric model that avoids any explicit assumptions on what form the hazard rate takes. This is in contrast to models such as Cox's proportional hazards model that focus on the hazard rate. We discuss the two different approaches in greater detail at the end of Chapter 4. Inference is done within the Bayesian formalism by numerically maximising the posterior density over function values. We construct the Laplace approximation of the hyperparameter posterior. The data likelihood terms corresponding to right, left, and interval censored and truncated observations are straightforward to compute.

This chapter is laid out as follows. In the first section we will adumbrate existing methods for analysing survival data, particularly interval censored data. In Section 3.3 we define a general transformation model from which several existing models can be derived as special cases (including our proposed GP regression model). Details of how parameter and hyperparameter inference are performed are given as well as some implementational details.

In Section 3.4 we compare our approach with some alternative approaches including a Weibull proportional hazards model (WPHM). We also compare our model against a model with a hazard rate given by $\lambda_0(\tau)\exp(f(\mathbf{x}))$, where $\lambda_0(\tau)$ is the base hazard rate, which is similar to the traditional Cox model but with an unspecified function of the covariates. A GP prior is assumed for this function which results in a model that allows for flexible non-linear covariate effects but imposes some additional structural assumptions on the form of the hazard rate. We finish in Section 3.5 with several simulation studies and comparisons of our model against some of the alternative models.

3.2 Existing methods

Here we give a brief and non-exhaustive overview of statistical methods for analysing survival data with one risk and independent censoring.

Non-parametric estimators

A common non-parametric estimator of the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958). To construct the estimator we reorder individuals such that $\tau_1 \leq \cdots \leq \tau_N$. The estimator is defined by

$$\hat{S}_{KM}(\tau) = \begin{cases} 1 & \text{if } \tau < \tau_1 \\ \prod_{i:\tau_i \leq \tau} \left(1 - \frac{D_i}{N_i}\right) & \text{otherwise,} \end{cases} \tag{3.1}$$

where D_i is the number of individuals for which the risk occurs at time τ_i and N_i is the number of individuals 'at risk' at τ_i. This includes all individuals still alive and uncensored up to and included time τ_i.

The Kaplan-Meier estimator can also be used to estimate the cumulative hazard rate since $\int_0^\tau \mathrm{d}s\, \pi(s) = -\log(S(\tau))$. An alternative estimator is the Nelson-Aalen (Nelson (1972) and Aalen (1978)) defined by

$$\hat{\Lambda}_{NL}(\tau) = \begin{cases} 0 & \text{if } \tau < \tau_1 \\ \sum_{i:\tau_i \leq \tau} \frac{D_i}{N_i} & \text{otherwise.} \end{cases} \tag{3.2}$$

The slope of $\hat{\Lambda}_{NL}$ gives an estimate of the hazard rate.

Semi-parametric and non-parametric models

Arguably the most popular method for analysing survival data is the Cox proportional hazards model (Cox, 1972). This approach assumes a semi-parametric hazard rate where there time dependence and the covariate effects factorise:

$$\pi_i(\tau) = \lambda_0(\tau) e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \quad \text{for } i = 1, \ldots, N. \tag{3.3}$$

Cox originally used a partial likelihood argument which avoids specification of the base hazard rate. This is equivalent to assuming Breslow's non-parametric estimator of the base hazard rate (originally given in the discussion section of Cox (1972)) and doing full likelihood estimation. Breslow's estimator for the cumulative base hazard is

$$\hat{\Lambda}_0(\tau) = \sum_{\tau_i \leq \tau} \frac{1}{\sum_{j \in R(\tau_i)} e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}_j}} \tag{3.4}$$

where $R(\tau_i)$ is the *risk set* containing all individuals who have not experienced any event up until time τ_i. There have been numerous extensions of Cox's model, many of which attempt to accommodate more complicated covariate effects by assuming $\pi(\tau) = \lambda_0(\tau)\exp(f(\mathbf{x}))$ where f is some function of the covariates. Generalised additive models assume $f(\mathbf{x}) = \boldsymbol{\beta}\cdot\mathbf{x} + \sum_{\mu=1}^{d} g_\mu(x_\mu)$ where g_μ are non-linear functions of the covariates (Fahrmeir and Kneib, 2011). See Martino et al. (2011) and Vanhatalo et al. (2013) for recent implementations of such models. Alternatively, a GP prior can be assumed for f as shown by Savitsky et al. (2011) and Joensuu et al. (2012). Viewed in this order these models seek to accommodate increasingly complicated covariate effects through more flexible and sophisticated functions of the covariates. We explore the GP model further in Section 3.4.3.

De Iorio et al. (2009) developed a flexible non-parametric model based on Dirichlet process priors that avoids the proportional hazards assumption.

Parametric models

Accelerated failure time (AFT) models approach the analysis of survival data in the spirit of traditional linear regression. The times are transformed so that they can take negative and positive values via $t = \log(\tau)$. These are related to covariates by

$$t_i = \boldsymbol{\beta}\cdot\mathbf{x}_i + \xi_i \quad \text{for } i = 1, \ldots, N \tag{3.5}$$

where $\boldsymbol{\beta}$ is a vector or regression coefficients. By choosing different distributions for the noise variable, ξ_i, one can recover different parametric regression models. For example choosing the extreme value distribution $p_\xi(s) = \exp(s - e^s)$ corresponds to a Weibull density over τ. See Klein and Moeschberger (2003, Section 2.6) for further details. See the textbook by Royston and Lambert (2011) for further examples of parametric survival models.

Frailty models

Frailty models (Vaupel et al., 1979) attempt to capture effects that 'missing' covariates may have on the hazard rates. This missing information manifests itself in the form of heterogeneity where individuals with identical covariates may nevertheless appear to have different hazard rates. This is dealt with by multiplying the hazard rate by a *frailty* factor w_i. If the hazard rate for individual i was originally $\pi_i^0(t)$ then it now becomes

$$\pi_i(t) = w_i \pi_i^0(t).$$

Since the individual frailty terms are unobserved some probability distribution is assumed and the frailty terms are subsequently integrated out. A common choice is the gamma distribution with the mean equal to one so as to eliminate redundancy in the overall scale of the hazard rate. See the textbook by Wienke (2011) for further details. A similar approach is taken to deal with competing risks which we shall discuss in Chapter 4.

Methods for interval censored data

There are three broad families of existing models for analysing interval censored data. Non-parametric estimators based on survival functions that are constant within disjoint intervals have been proposed by Peto (1973) and Turnbull (1976). Secondly, parametric models assume a specific parametric event time density. Popular choices are the Weibull, exponential or log-Gaussian densities for example. The advantage of parametric models is that expressions for the survival function can be obtained in closed form and hence the exact likelihood can be constructed for right, left or interval censored observations. Covariate effects can be included via a link function which specifies that some parameter of the probability density is a function of the covariates. Numerical methods can be used to infer unknown parameter values. See Lindsey (1998) for a discussion and comparison of several parametric models. Odell et al. (1992), Rabinowitz et al. (1995) and Komárek and Lesaffre (2009) consider Weibull accelerated

failure time models. Sparling et al. (2006) present a family of parametric models that can handle time dependent covariates.

Finally, there are semi-parametric models, of which most are adaptations of the Cox proportional hazards model. The partial likelihood argument used by Cox cannot be used in the presence of interval censoring. However, the full likelihood can be written in terms of the event time density and survival functions and this can be numerically optimised with respect to any model parameters (Finkelstein, 1986). Markov Chain Monte Carlo methods have been used by Sinha et al. (1999) in a Bayesian discretised Cox model and by Satten (1996) in a proportional hazards model. The EM algorithm has been used by Goggins et al. (1998) and Goetghebeur and Ryan (2000) to infer parameters in proportional hazards models. Several authors use smoothing techniques to model the base hazard rate (Betensky et al., 2002) or the event time density (Zhang and Davidian, 2008). Kooperberg and Clarkson (1997) and Zhang et al. (2010) used splines to model a smooth hazard rate. Another strategy is to impute the event times (Law and Brookmeyer, 1992) by taking the midpoint or the end of the interval for instance (Pan, 2000), and then applying standard methods to the imputed event times.

3.3 The Gaussian process regression model

We firstly define a general non-linear transformation model from which several existing models can be recovered under different assumptions. This will serve as a starting point for the GP regression model and offer an intuitive way to compare it to existing approaches. We then provide details of how to infer parameters and make predictions for new individuals.

3.3.1 General non-linear transformation model

A general transformation model assumes that

$$\phi(\tau_i) = f(\mathbf{x}_i) + \xi_i \quad \text{for } i = 1, \ldots, N \tag{3.6}$$

where ϕ is a monotonically increasing transformation of the event times, $f(\mathbf{x}_i)$ is some function of the covariates, and ξ_i is a noise random variable with a probability density function p_ξ.

Under different assumptions of ϕ, f and $p(\xi)$ several existing models, including Gaussian process models, can be derived as special cases of (3.6). For example, linear transformation models assume ϕ is unspecified and $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$. Various procedures for estimating the regression parameters in such models have been proposed by Cheng et al. (1995), Fine et al.

(1998) and Chen et al. (2002). Recently Lu and Li (2008) considered the case where $f(\mathbf{x})$ is an unspecified smooth function and proposed a boosting estimation method based on the marginal likelihood.

If we pick $p_\xi(s) = \exp(s - e^s)$ and $\phi(\tau) = \log \Lambda_0(\tau)$ we recover models with a hazard rate similar to Cox's model. To see this we write $\xi_i = \log \Lambda_0(\tau) - f(\mathbf{x}_i)$ and derive the event time density

$$\begin{aligned} p(\tau_i) &= p_\xi \left(\log \Lambda_0(\tau_i) - f(\mathbf{x}_i)\right) \frac{\mathrm{d}}{\mathrm{d}\tau} \left(\log \Lambda_0(\tau_i) - f(\mathbf{x}_i)\right) \\ &= \lambda_0(\tau_i) e^{-f(\mathbf{x}_i)} \exp(-\Lambda_0(\tau_i) e^{-f(\mathbf{x}_i)}). \end{aligned} \tag{3.7}$$

We can readily verify that this corresponds to a hazard rate similar to Cox's model by substituting $\pi_i(\tau) = \lambda_0(\tau) \exp(-f(\mathbf{x}_i))$ into (2.32). When $f(\mathbf{x}) = -\boldsymbol{\beta} \cdot \mathbf{x}$ we recover Cox's original proportional hazards model. Frailty models can be retrieved by assuming $f(x) = -\boldsymbol{\beta} \cdot \mathbf{x} + w$ where w is a frailty term. Generalised additive models assume $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x} + \sum_{\mu=1}^{d} g_\mu(x_\mu)$ where g_μ are non-linear functions of the covariates as discussed above. When a GP prior is assumed over $f(\mathbf{x})$ we recover the model used by Joensuu et al. (2012). For completeness we note that accelerated failure time models can be recovered by assuming $\phi(\tau) = \log(\tau)$ and $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$. Assuming different distributions for ξ results in a wide variety of accelerated failure time models.

3.3.2 Gaussian process prior for the latent function values

From now on we let $t = \phi(\tau)$ denote the transformed event times. We could choose the traditional $t = \log(\tau)$ but instead we choose

$$t = \phi(\tau) = \log(e^{\tau/\gamma} - 1). \tag{3.8}$$

This transformation has some desirable features. Provided $\gamma < \min_i(\tau_i)$ then the transformation is effectively linear. A log transformation will be non-linear and this will become particularly apparent for large $\phi(\tau)$. We may have two large values of τ that once transformed are rather similar to each each other. This may make it difficult for the model to make accurate inferences for large values of τ. Since we will be assuming a Gaussian noise model the uncertainty associated with large event times will the same as for short event times but with a non-linear transformation this is not desirable. Therefore (3.8) is preferable. The distortion due to the non-linear component of the transformation (when $\tau < \gamma$) becomes ap-

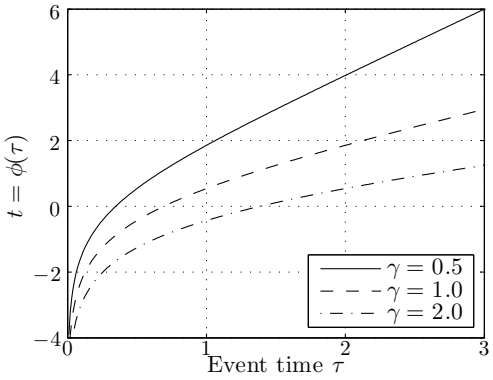


Figure 3.1: Plot of the time transformation $\phi(\tau)$ that was used in the GP implementation. Note that for values of $\tau > \gamma$ the transformation is approximately linear. By adjusting the value of γ such that it is less than the earliest observed event time we effectively end up with a linear mapping. The effect of the transformation can be seen when negative values of $t = \phi(\tau)$ are predicted since they are 'squashed' into the positive half of the real line.

parent only during predictions. When t takes negative values they are 'squashed' towards the positive half of the real line. The transformation is plotted for various values of γ in Figure 3.1.

The transformation of the output variables in Gaussian process regression has been explored by Snelson et al. (2004). They examine a variety of parameterised monotonic transformations and regard any transformation parameters as hyperparameters to learn during training. Their procedure infers the most appropriate transformation such that the transformed outputs can be modelled using a Gaussian process. It may be useful to apply this method in future work.

To construct a Gaussian process (GP) model we assume a GP prior over the latent function values $f(\mathbf{x}_i)$:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta})}}{(2\pi)^{N/2}|\mathbf{K}|^{1/2}}. \tag{3.9}$$

In this Chapter we have used the squared exponential kernel and the linear kernel defined in (2.49). For the noise variable in (3.6) we pick $p(\xi) = \mathcal{N}(0, \beta^2)$ and it follows that the event time density is

$$p(t_i|f(\mathbf{x}_i)) = \mathcal{N}(f(\mathbf{x}_i), \beta^2). \tag{3.10}$$

This has a convenient form since the conditional event time density has a simple form with all

of the non linear covariate effects captured by $p(\mathbf{f}|\mathbf{X})$. From this we can derive the survival function and hazard rate using (2.27, 2.28):

$$S(\tau) = \int_{\tau}^{\infty} \mathrm{d}s\, p(s|f_i) \quad \text{and} \quad \pi_i(\tau) = \frac{p(\tau|f_i)}{\int_{\tau}^{\infty} \mathrm{d}s\, p(s|f_i)}. \tag{3.11}$$

3.3.3 Inference of latent function values and hyperparameters

For the present section we will consider only right censoring. Interval censoring will be considered in Section 3.3.6. We need to infer the values of N latent function values $\mathbf{f}$ and the values of any hyperparameters that are used in the kernel function. We infer the latent function values using Bayes' theorem (2.34):

$$p(\mathbf{f}|\mathbf{X}, D, \boldsymbol{\theta}) = \frac{p(D|\mathbf{f}, \boldsymbol{\theta})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{\int \mathrm{d}\mathbf{f}'\, p(D|\mathbf{f}', \boldsymbol{\theta})p(\mathbf{f}'|\mathbf{X}, \boldsymbol{\theta})} \tag{3.12}$$

with $D = \{(\tau_1, \Delta_1), \ldots, (\tau_N, \Delta_N)\}$ and $p(D|\mathbf{f}, \boldsymbol{\theta}) = \prod_{i=1}^{N} P_i(t_i, \Delta_i|f_i)$ where $P_i(t_i, \Delta_i|f_i)$ depends on what type of censoring or truncation has occurred and is given by (2.30). We determine the maximum a posteriori (MAP) solution by numerically minimising the negative log likelihood:

$$\begin{aligned}\mathcal{L}(\mathbf{f}) &= -\frac{1}{N}\log p(\mathbf{f}|\mathbf{X}, D, \boldsymbol{\theta}) \\ &= -\frac{1}{N}\sum_{i:\Delta_i=1}\log p(t_i|f_i) - \frac{1}{N}\sum_{i:\Delta_i=0}\log S(t_i|f_i) \\ &\qquad + \frac{1}{2N}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}) + \frac{1}{2N}\log|\mathbf{K}|. \end{aligned} \tag{3.13}$$

Numerical optimisation is performed using a gradient based optimiser in Matlab. Partial derivatives are given in Appendix A.1. Hyperparameters are determined by optimising the Laplace approximation of the marginal likelihood $\int \mathrm{d}\mathbf{f}'\, p(D|\mathbf{f}', \boldsymbol{\theta})p(\mathbf{f}'|\mathbf{X}, \boldsymbol{\theta})$ (from (2.44)):

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|\mathbf{W} + \mathbf{K}^{-1}| \tag{3.14}$$

where the diagonal matrix is defined by $\mathbf{W}_{ii} = -\frac{\partial^2}{\partial f_i^2}\log p(D|\mathbf{f}, \boldsymbol{\theta})$ (see Appendix A.1) and $\hat{\mathbf{f}} = \min_{\mathbf{f}} \mathcal{L}(\mathbf{f})$ is obtained by minimising (3.13). Note that each evaluation of the hyperparameter posterior requires finding $\hat{\mathbf{f}}$. Further details are given in Section 3.3.7.

3.3.4 Posterior properness

In the case where there is no censoring then the posterior (3.12) is the same as that obtained in standard GP regression with noisy outputs. This posterior is a multivariate Gaussian density which is proper (that is, the integral of the posterior is finite):

$$\begin{aligned}\int \mathrm{d}\mathbf{f}\, p(D|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) &= \int \mathrm{d}\mathbf{f} \prod_{i=1}^{N} p(t_i|f_i)p(\mathbf{f}|\mathbf{X}) \\ &= \frac{e^{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\eta})\cdot\tilde{\mathbf{K}}(\mathbf{t}-\boldsymbol{\eta})}}{(2\pi)^{N/2}|\tilde{\mathbf{K}}|^{1/2}} \end{aligned} \tag{3.15}$$

with $[\mathbf{t}]_i = t_i$ and $\tilde{\mathbf{K}} = \mathbf{K} + \beta^2\mathbf{I}$. In the case where some individuals are right censored then then some of the Gaussian terms $p(t_i|f_i)$ will be replaced with cumulative Gaussian terms $S(t_i|f_i)$. This effectively *truncates* regions of the posterior that are inconsistent with survival outcomes[1]. The survival function is bounded above and below $0 \leq S(t|f) \leq 1$, and we can use this to show that the posterior is proper when we have right censoring:

$$0 \leq \int \mathrm{d}\mathbf{f} \prod_{i:\Delta_i=1}^{N} p(t_i|f_i) \prod_{i:\Delta_i=0}^{N} S(t_i|f_i)p(\mathbf{f}|\mathbf{X}) \leq \int \mathrm{d}\mathbf{f} \prod_{i:\Delta_i=1}^{N} p(t_i|f_i)p(\mathbf{f}|\mathbf{X}) < \infty. \tag{3.16}$$

Note that this is true for any permissible choice of kernel function k.

3.3.5 Predictions, hazard rates and survival curves

Having observed data D and trained a GP regression model (by inferring latent function values $\mathbf{f}$ and hyperparameters $\boldsymbol{\theta}$) we may wish to predict the event time τ^* for a new individual with covariates $\mathbf{x}^*$. The predictive distribution for a test output $\mathbf{f}^*$ corresponding to a test input $\mathbf{x}^*$ is Gaussian with mean and variance

$$\hat{\mu} = \mathbf{k}^* \cdot \mathbf{K}^{-1}\hat{\mathbf{f}} \tag{3.17}$$

$$\hat{\kappa} = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^* \cdot (\mathbf{K} + \mathbf{W}^{-1})^{-1}\mathbf{k}^*. \tag{3.18}$$

[1]The posterior obtained in Gaussian process classification has a similar structure (Rasmussen and Williams, 2006, Section 3.3).

These expression are similar to the standard GP predictive mean (2.51) and variance (2.52) except in this case we include additional variance[2] due to the uncertainty in $\hat{\mathbf{f}}$.

The corresponding density for (the noisy prediction) t^* is $\mathcal{N}(\hat{\mu}, \hat{\kappa} + \beta^2)$. Finally, we need to transform back to the original time variable τ^*:

$$p(\tau^*|\mathbf{x}^*, \mathbf{X}, D) = \frac{e^{-\frac{1}{2(\hat{\kappa}+\beta^2)}(\log(e^{\tau^*/\gamma}-1)-\hat{\mu})^2}}{(2\pi(\hat{\kappa}+\beta^2))^{1/2}} \frac{e^{\tau^*/\gamma}}{\gamma(e^{\tau^*/\gamma}-1)}. \tag{3.19}$$

Once the predictive event time density has been obtained we can compute the primary hazard rate if desired:

$$\pi(\tau^*|\mathbf{x}^*, \mathbf{X}, D) = \frac{p(\tau^*|\mathbf{x}^*, \mathbf{X}, D)}{S(\tau^*)} \tag{3.20}$$

where the survival function is

$$S(\tau^*|\mathbf{x}^*, \mathbf{X}, D) = \int_{\tau^*}^{\infty} \mathrm{d}s\, p(s|\mathbf{x}^*, \mathbf{X}, D). \tag{3.21}$$

It may also be desirable to make a specific prediction of when the event will occur. This can be done by numerically computing the mean of the event time density:

$$\langle \tau^* \rangle = \int_0^{\infty} \mathrm{d}s\, sp(s|\mathbf{x}^*, \mathbf{X}, D). \tag{3.22}$$

The variance $\left\langle (\tau^*)^2 \right\rangle - \langle \tau^* \rangle^2$ can also be computed as gives us a measure of uncertainty regarding our prediction.

3.3.6 Application to interval censored data

We also implement a model that accommodates interval censored observations. We assume that all of the observations are either interval censored or right censored but it would be straightforward to relax this and include additional types or non-censored, censored or truncated observations. For an individual who is interval censored we observe upper and lower times that define an interval[3] (t_i^l, t_i^u) and we have $\mathrm{prob}(t_i^l, t_i^u, \Delta_i = 1|f_i) = S(t_i^l) - S(t_i^u)$.

[2] See Section 3.4.2 of Rasmussen and Williams (2006).

[3] Note that we are working with the transformed event times $t = \phi(\tau)$ defined by (3.8).

Taking the negative log of the posterior (3.12) and ignoring terms independent of $\mathbf{f}$ we get

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log[S(t_i^l|f_i) - S(t_i^u|f_i)] - \frac{1}{N}\sum_{i:\Delta_i=0}\log S(t_i|f_i) - \frac{1}{N}\log p(\mathbf{f}|\mathbf{X}). \tag{3.23}$$

As above, we find $\hat{\mathbf{f}}$ by numerically minimising the negative log likelihood. Hyperparameters are determined using the Laplace approximation of the marginal likelihood:

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|\mathbf{W} + \mathbf{K}^{-1}| \tag{3.24}$$

where the diagonal matrix $\mathbf{W}$ is defined by $\mathbf{W}_{ii} = -\frac{\partial^2}{\partial f_i^2}\log p(D|\mathbf{f}, \boldsymbol{\theta})$ and $\hat{\mathbf{f}} = \min_{\mathbf{f}} \mathcal{L}(\mathbf{f})$. First and second order partial derivatives are given in Appendix A.2. Once we have obtained $\hat{\mathbf{f}}$ and the hyperparameters predictions for new individuals proceed in exactly the same way as described above in Section 3.3.5.

3.3.7 Numerical implementation

A number of numerical issues arise during the implementation of the above models. Numerical instability can occur when computing the negative log likelihood function (3.13). The problematic terms are the hazard rates $\pi(t|f) = p(t|f)/S(t|f)$ where $p(t|f)$ is a Gaussian density and $S(t|f)$ is the corresponding survival function. The hazard rates do not appear in (3.13) but the same terms do occur in the partial derivatives (see (A.4) and (A.7)). This quantity can be written in terms of the *complementary error function*

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \mathrm{d}s\, e^{-s^2}. \tag{3.25}$$

If we define $h = (t - f)/\beta\sqrt{2}$ then

$$\log S(t|f) = \log\left(\tfrac{1}{2}\operatorname{erfc}(h)\right) \tag{3.26}$$

$$\frac{\partial}{\partial f}\log S(t|f) = \frac{2}{\sqrt{2\pi}\beta}\frac{e^{-h^2}}{\operatorname{erfc}(h)} \tag{3.27}$$

$$\frac{\partial^2}{\partial f^2}\log S(t|f) = -\left(\frac{2}{\sqrt{2\pi}\beta}\frac{e^{-h^2}}{\operatorname{erfc}(h)}\right)^2 + \frac{\sqrt{2}h}{\beta}\left(\frac{2}{\sqrt{2\pi}\beta}\frac{e^{-h^2}}{\operatorname{erfc}(h)}\right). \tag{3.28}$$

The hazard rate is also given by (3.27). For large h the quantity $e^{-h^2}/\text{erfc}(h)$ becomes numerically unstable since both numerator and denominator tend towards zero. This is solved by using the asymptotic expansion of the complementary error function (Menzel, 1960):

$$\text{erfc}(h) = \frac{e^{-h^2}}{h\sqrt{\pi}}\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right] \quad \text{for } h \gg 0. \tag{3.29}$$

This results in the following approximations which are numerically stable:[4]

$$\log S(t|f) = -h^2 - \log h - \log 2\sqrt{\pi} + \log\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right] \tag{3.30}$$

$$\frac{\partial}{\partial f}\log S(t|f) = \frac{h\sqrt{2}}{\beta}\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right]^{-1}. \tag{3.31}$$

The approximation (3.27) can be substituted directly into (3.28) to obtain an approximation for the second order partial derivative. Note that the hazard rate is approximately linear for large h which is consistent with Figure 3.4 (d).

3.4 Comparison to hazard rate models

In this section we will compare the performance of our GP regression model to models that assume a more traditional hazard rate of the form

$$\pi(\tau) = \lambda_0(\tau)\exp(f(\mathbf{x})) \tag{3.32}$$

for some function $f(\mathbf{x})$ and base hazard rate $\lambda_0(\tau)$. We will examine the standard Cox model, a Weibull proportional hazards model (WPHM) and the model of Joensuu et al. (2012) which assumes a GP prior over the function values $\mathbf{f}$ in (3.32).

It is not obvious how to compare the performance of different models. From the point of view of GP regression it is natural to think in terms of inferring a function between the outputs (event times) and inputs (covariates). There are different quantities that measure the 'goodness of fit' such as computing the mean square error (MSE) between the inferred function and the reported event times. An even better approach is to split the dataset into a training and validation set and compute the MSE between the predicted time-to-event and the reported time-to-event in both sets. This approach tests the models' ability to generalise

[4]Using the approximations for $h > 20$ gives acceptable performance in Matlab.

to unseen individuals since good performance on the validation set indicates that the model has indeed extracted genuine structure from the training set. A hallmark of overfitting is when very good performance is attained on the training set but with poor performance on the validation set. This occurs when the model fits to noise or detects spurious relationships that do not occur in the validation data.

However, inferring a function and using it to make predictions is unsuitable with Cox's original proportional hazards model since the corresponding event time density is unnormalised and the mean of the predictive density is not always defined. We explain this in more detail below in Section 3.4. This limitation motivates us to use the WPHM since the event time density is correctly normalised and the model can be used to make well-defined predictions by computing the mean of the event time density.

Nevertheless, we can still compare survival curves from a Cox model and our GP model. The individuals can be ranked in order of $\boldsymbol{\beta}\cdot\mathbf{x}_i$ where negative values indicate a longer survival time and positive values correspond to shorter survival. Once the individuals are ranked they can be split into 'high' and 'low' risk groups (we can also examine tertiles or quartiles) and Kaplan-Meier survival curves for each group can be plotted. If the model fits well then we would expect to see a difference between the survival curves. This procedure can also be performed after splitting the cohort into training and validation groups, training the model on the training group, and then using the model to generate survival curves for individuals belonging to the validation group. This approach allows us to check if the model is overfitting (which would be characterised by poor generalisation to unseen individuals).

3.4.1 The Cox proportional hazards model

The Cox model (with Breslow's estimator of the base hazard rate) is ill suited to predicting an actual event time (and better suited for establishing associations covariates and survival outcomes). This is due to the fact that the event time density is not normalised and consequently may not have a well-defined mean or variance. Breslow's estimate of the cumulative base hazard rate is

$$\hat{\Lambda}_0(\tau) = \sum_{\tau_i \leq \tau} \frac{1}{\sum_{j \in R(\tau_i)} e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}_j}} \tag{3.33}$$

and was originally presented in the discussion section of Cox (1972). The *risk group* $R(\tau_j)$ is the set of all individuals who are still 'at risk' (i.e. still alive) at time τ_j. The regression parameters $\hat{\boldsymbol{\beta}}$ are maximum likelihood estimators obtained from the partial likelihood method.

As noted in Klein and Moeschberger (2003, Section 8.3) Breslow's estimator is the maximum likelihood estimator of the full likelihood (for fixed $\boldsymbol{\beta}$):

$$p(D|\boldsymbol{\beta}, \lambda_0) = \prod_{i=1}^{N} [\lambda_0(\tau_i) e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}]^{1-\Delta_i} \exp(-\Lambda_0(\tau) e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}). \tag{3.34}$$

Substituting (3.33) into (3.34) and maximising with respect to $\boldsymbol{\beta}$ yields the same estimate for $\boldsymbol{\beta}$ as obtained from the partial likelihood argument (Coolen and Holmberg, 2014, Section 8.1).

Once $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(\tau)$ have been estimated from the data the event time density corresponding to an individual with covariates $\mathbf{x}^*$ is obtained from (2.32):

$$p(\tau|\mathbf{x}^*, \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0) = \hat{\lambda}_0(\tau_i) e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*} \exp(-\hat{\Lambda}_0(\tau) e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}). \tag{3.35}$$

Ideally it whould be possible to use this density to make predictions with corresponding error bars by computing the mean $\langle\tau\rangle$ and variance $\langle\tau^2\rangle - \langle\tau\rangle^2$. However, the density (3.35) is not normalised when Breslow's estimator is used. To see this we integrate

$$\int_0^\infty \mathrm{d}s\, p(s|\mathbf{x}^*, \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0) = 1 - \exp(-\hat{\Lambda}_0(s) e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*})\Big|_{s=\infty}. \tag{3.36}$$

Correct normalisation requires

$$\lim_{\tau\to\infty} \Lambda(\tau) = \infty, \tag{3.37}$$

a condition that is not met by Breslow's estimator since the largest value (3.33) can take occurs after the largest observed time, $\max_i(\tau_i)$, and which we denote by $\hat{\Lambda}_0^\infty$

$$\hat{\Lambda}_0^\infty = \sum_{\tau_i} \frac{1}{\sum_{j\in R(\tau_i)} e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}_j}} < \infty. \tag{3.38}$$

Nevertheless survival curves can be generated according to

$$S(\tau|\mathbf{x}^*, \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0) = \exp(-\hat{\Lambda}_0(\tau) e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}). \tag{3.39}$$

However the survival function will never reach zero, even for infinitely large time, which implies a finite probability of immortality. This is consistent with an unnormalised probability density (3.36) which means there is a finite probability the event will never occur. In fact,

the survival curve (3.39) and the integrated base hazard rate (3.33) are constant after the last observed event. While Cox's model may provide a useful description of what occurs within the timespan of observed data the incorrect normalisation is clearly undesirable and is unsuitable for predicting the event time for new individuals. As such, we examine the WPHM which avoids this problem.

3.4.2 The Weibull proportional hazards model

For the purposes of making predictions we will implement a Weibull proportional hazards model (WPHM) but without the problematic Breslow's estimator. We choose a Weibull base hazard rate

$$\lambda_0(\tau) = (\nu/\rho)(\tau/\rho)^{\nu-1} \tag{3.40}$$

where $\rho > 0$ is a scale parameter and $\nu > 0$ is a shape parameter. It follows that the cumulative base hazard rate is $\Lambda_0(\tau) = (\tau/\rho)^\nu$. Note that the normalisation condition (3.37) is satisfied. The hazard rate for individual i is

$$\pi_i(\tau|\mathbf{x}_i, \nu, \rho, \boldsymbol{\beta}) = \lambda_0(\tau)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}. \tag{3.41}$$

Using Bayes' theorem the posterior over parameters is $p(\boldsymbol{\beta}, \rho, \nu|D) \propto p(D|\boldsymbol{\beta}, \rho, \nu)p(\boldsymbol{\beta})p(\rho)p(\nu)$. The data likelihood is

$$p(D|\boldsymbol{\beta}, \nu, \rho) = \prod_{i=1}^{N} [\lambda_0(\tau_i)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}]^{\Delta_i} \exp(-\Lambda_0(\tau)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}). \tag{3.42}$$

We can then define the negative log likelihood as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \rho, \nu) &= -\frac{1}{N}\log p(\boldsymbol{\beta}, \rho, \nu|D) \\ &= -\frac{1}{N}\sum_{i:\Delta_i=1}[\log\lambda_0(\tau_i) + \boldsymbol{\beta}\cdot\mathbf{x}_i] + \frac{1}{N}\sum_{i=1}^{N}\Lambda_0(\tau_i)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \\ &\quad - \frac{1}{N}\log p(\boldsymbol{\beta}) - \frac{1}{N}\log p(\rho) - \frac{1}{N}\log p(\nu)\end{aligned} \tag{3.43}$$

where $\log\lambda_0(\tau) = \log\nu - \log\rho + (\nu-1)\log(\tau/\rho)$. We assume $p(\boldsymbol{\beta})$, $p(\rho)$ and $p(\nu)$ are constant (and therefore improper) priors[5]. The optimal values of the parameters are given by

[5]The posterior is proper however.

numerically solving

$$\{\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\nu}\} = \text{argmax}_{\boldsymbol{\beta},\rho,\nu}\mathcal{L}(\boldsymbol{\beta}, \rho, \nu). \tag{3.44}$$

The gradient based 'fminsearch' optimisation function is used in Matlab. Partial derivatives can be found in Appendix C.

Finally, error bars for β_μ can be obtained from $\sqrt{(N\mathbf{H})^{-1}_{\mu\mu}}$ where $\mathbf{H}$ is the matrix of second order partial derivatives given in Appendix C. This gives the standard deviation of that parameter under a Gaussian approximation of the posterior. Error bars for ρ and ν are not defined under a Gaussian approximation since by definition both parameters are non-negative. In the implementation both parameters are reparameterised such that they take real values via

$$\rho \to \log(1 + \rho_{LB} + e^{\rho}) \quad \text{and} \quad \nu \to \log(1 + \nu_{LB} + e^{\nu}), \tag{3.45}$$

where $\rho_{LB} \geq 0$ and $\nu_{LB} \geq 0$ are lower bounds on the respective parameters. Predictions can be made by computing the mean (and variance) of the event time density corresponding to a new individual with covariates $\mathbf{x}^*$

$$\langle\tau\rangle = \int_0^\infty \mathrm{d}s\, s\hat{\lambda}_0(s)e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}\exp(-\hat{\Lambda}_0(s)e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}). \tag{3.46}$$

The hazard rate and survival function are respectively given by

$$\pi(\tau|\mathbf{x}^*, \hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\nu}) = (\hat{\nu}/\hat{\rho})(\tau/\hat{\rho})^{\hat{\nu}-1}e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*} \tag{3.47}$$

$$S(\tau|\mathbf{x}^*, \hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\nu}) = e^{(\tau/\hat{\rho})^{\hat{\nu}}e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}}. \tag{3.48}$$

3.4.3 The Joensuu Gaussian process hazard rate model

Finally, we will compare all of the above models to a second type of GP model that assumes a hazard rate $\pi_i(\tau|\mathbf{x}_i) = \lambda_0(\tau)\exp(f(\mathbf{x}_i))$ with $f(\mathbf{x})$ an unspecified function of the covariates. A GP prior is assumed for this function $p(\mathbf{f}|\mathbf{X})$. Such models were used by Joensuu et al. (2012) with a piecewise log-constant base hazard rate. The Laplace approximation of the likelihood was constructed in order to estimate the marginal likelihood and infer hyperparameters. The same model is discussed in Savitsky et al. (2011). We will use a base hazard rate corresponding to the Weibull distribution $\lambda_0(\tau) = \nu\tau^{\nu-1}$ with $\nu > 0$, which implies $\Lambda_0(\tau) = \tau^\nu$. The likelihood contribution for individual i is written in terms of the hazard rate

$$P_i(\tau_i, \Delta_i|f_i) = \pi_i(\tau)^{\Delta_i}e^{-\int_0^\tau \mathrm{d}s\,\pi_i(s)}. \tag{3.49}$$

The posterior (3.12) is also valid here but the data likelihood terms will be different. The negative log posterior is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\log p(\mathbf{f}|D) \tag{3.50}$$

$$= -\frac{1}{N}\sum_{i:\Delta_i=1}\Big[\log\lambda_0(\tau_i) + f(\mathbf{x}_i)\Big] + \frac{1}{N}\sum_{i=1}^{N}\Lambda_0(\tau_i)e^{f(\mathbf{x}_i)} + \frac{1}{2N}\mathbf{f}\cdot\mathbf{K}^{-1}\mathbf{f} + \frac{1}{2N}\log|\mathbf{K}| + \frac{1}{2}\log 2\pi. \tag{3.51}$$

Using (2.44) to construct the Laplace approximation of the marginal likelihood the negative log hyperparameter posterior is

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|\mathbf{W} + \mathbf{K}^{-1}| \tag{3.52}$$

with $\mathbf{W}_{ii} = -\frac{\partial^2}{\partial f_i^2}\log p(D|\mathbf{f})$ and $\hat{\mathbf{f}} = \min_{\mathbf{f}}\mathcal{L}(\mathbf{f})$. See Appendix A.3 for first and second order partial derivatives. The predictive density over f^* is Gaussian with mean and variance given by (3.17, 3.18). The event time density is

$$p(\tau|f(\mathbf{x}_i)) = \lambda_0(\tau)e^{f(\mathbf{x}_i)}e^{-\Lambda_0(\tau)e^{f(\mathbf{x}_i)}} \tag{3.53}$$

which can be used to obtain the predictive distribution over the event time:

$$p(\tau^*|\mathbf{x}^*, \mathbf{X}, D) = \int \mathrm{d}f^*\, \lambda_0(\tau^*)e^{f^*}e^{-\Lambda_0(\tau^*)e^{f^*}}\frac{e^{-\frac{1}{2\hat{\kappa}^2}(f^*-\hat{\mu})^2}}{(2\pi\hat{\kappa}^2)^{1/2}}. \tag{3.54}$$

This expression was found to be problematic in that $\int \mathrm{d}\tau^*\, p(\tau^*|\mathbf{x}^*, \mathbf{X}, D) \neq 1$. This is an example of non-commuting limits since if we first integrate the integrand in (3.54) with respect to τ^* and then f^* we obtain a value of one. A rough explanation of why this occurs is that negative values of f^* correspond to a protective effect. That is, the event time density places more probability mass away from the origin. As the value of f^* decreases more probability mass is placed further and further away from the origin. In the limit $f^* \to -\infty$ then $\mathrm{prob}(\tau^* = \infty) \to 1$.

We have not produced a more formal examination of this issue but experience suggests that numerical computation of the mean and variance of (3.54) is infeasible. Consequently, the predictive mean $\langle\tau^*\rangle$ and variance $\langle(\tau^*)^2\rangle - \langle\tau^*\rangle^2$ will be computed numerically from

(3.53). Note that this will underestimate the uncertainty since we are not taking into account the uncertainty in f^*.

3.4.4 Application of the Joensuu model to interval censored data

Since we have developed our GP model to include interval censored observations we extend the Joensuu model to incorporate interval censoring also. The negative log likelihood is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log S(\tau_i|f_i) - \frac{1}{N}\sum_{i:\Delta_i=0}^{N}\log[S(\tau_i^l|f_i) - S(\tau_i^u|f_i)] + \frac{1}{2N}\mathbf{f}\cdot\mathbf{K}^{-1}\mathbf{f} + \frac{1}{2N}\log|\mathbf{K}| + \frac{1}{2}\log 2\pi \tag{3.55}$$

which is numerically minimised with respect to $\mathbf{f}$. The survival function is $S(\tau|f_i) = \exp(-\Lambda_0(\tau)e^{f_i})$. As before, we construct the Laplace approximation of the marginal posterior to obtain the negative log hyperparameter posterior

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|N\mathbf{H}| \tag{3.56}$$

where $\mathbf{H}$ is the matrix of second order partial derivatives (see Appendix A.4 for first and second order partial derivatives) and $\hat{\mathbf{f}} = \min_{\mathbf{f}}\mathcal{L}(\mathbf{f})$.

Numerical implementation

Some numerical issues arise in the computation of $\log(S(\tau_i^l) - S(\tau_i^u))$ while we search the parameter space for an optimal solution (and when we compute the partial derivatives (A.20) and (A.22)). This is because $S(\tau) = \exp(-\Lambda_0(\tau)e^f)$ can take extremely small values and unlike the right censored case the log does not cancel the exponentials because of the sum. In this case we write

$$\log(e^{-x_1} - e^{-x_2}) = \log\{e^{-x_1}(1 - e^{x_1-x_2})\} \tag{3.57}$$

$$= \begin{cases} -x_1 + \log(1 - e^{x_1-x_2}) & \text{when } -C \le x_1 - x_2 \\ -x_1 - e^{x_1-x_2} - \frac{1}{2}e^{2(x_1-x_2)} & \text{when } x_1 - x_2 < -C. \end{cases} \tag{3.58}$$

Note that $x_2 \geq x_1$ in the context of this model. The constant C is a cutoff that depends on the numerical accuracy of implementation[6]. Furthermore, a similar problem occurs with the computation of first and second order gradients. A similar trick can rectify the problem. The offending terms from the gradient in Appendix A.4 are (A.20) and (A.22) and can be rewritten as

$$\frac{-x_1 e^{-x_1} + x_2 e^{-x_2}}{e^{-x_1} - e^{-x_2}} = \frac{-x_1 + x_2 e^{x_1 - x_2}}{1 - e^{x_1 - x_2}}, \tag{3.59}$$

and the second order derivatives are given by

$$-\delta_{ij}\left(\frac{-x_1 + x_2 e^{x_1-x_2}}{1 - e^{x_1-x_2}}\right)^2 + \delta_{ij}\frac{-x_1 + x_1^2 + (x_2 - x_2^2)e^{x_1-x_2}}{1 - e^{x_1-x_2}}. \tag{3.60}$$

3.5 Results

In this section we present results from simulated data and compare our model to the Cox's proportional hazards model, the WPHM and the GP model of Joensuu.

3.5.1 Generation of simulated survival data

Simulated data with a specified hazard rate

For the purposes of comparison we wish to generate simulated data according to the WPHM model outlined in Section 3.4.2. We begin by choosing values of $\boldsymbol{\beta}$, ρ, ν manually. Covariate vectors $\mathbf{x}_i$ are generated from a uniform distribution on a finite region of the covariate space. Event times are generated using the inverse of the cumulative distribution. From (2.31) this is

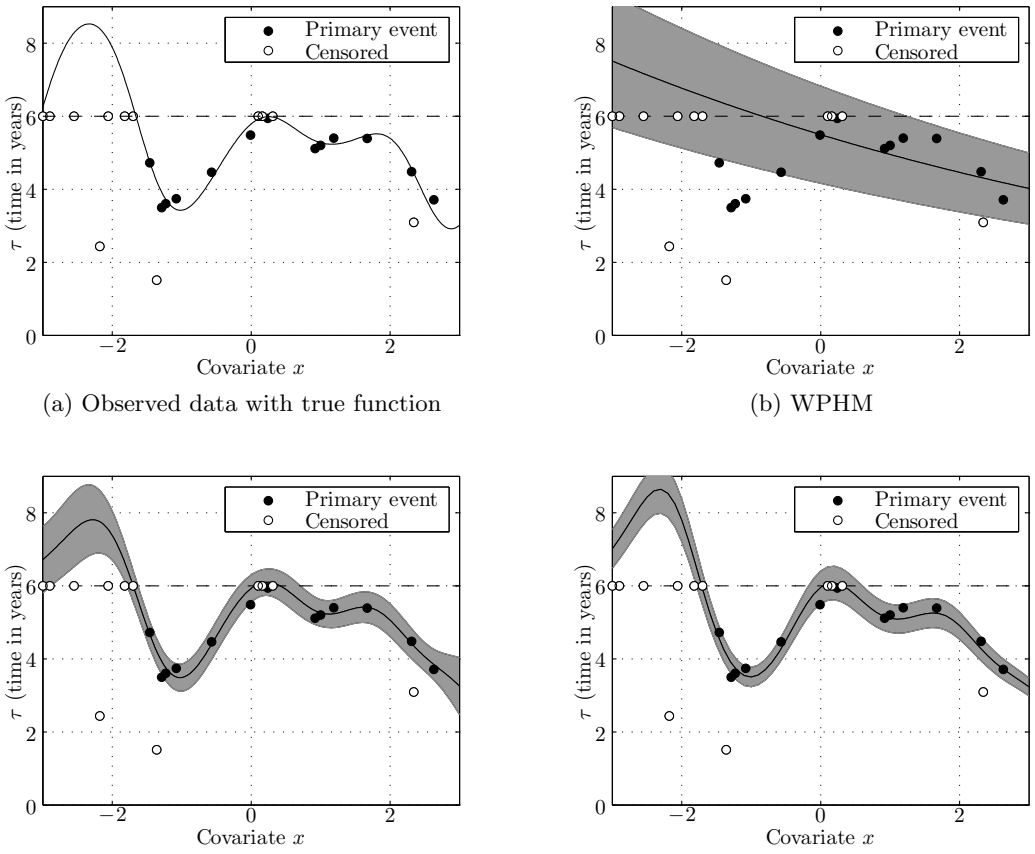
$$C_i(\tau) = 1 - e^{-\Lambda_0(\tau)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}}. \tag{3.61}$$

Random numbers $z \in [0, 1]$ are generated from a uniform density. An event time corresponding to $\mathbf{x}_i$ is

$$\tau_i = \rho\left(-e^{-\boldsymbol{\beta}\cdot\mathbf{x}_i}\log(1-z)\right)^{1/\nu}. \tag{3.62}$$

Finally independent censoring is simulated by randomly selecting a subset of the individuals and generating a random number from a uniform distribution defined on the interval $[0, \tau_i)$ which is then recorded as the time of censoring.

[6] In Matlab $C = 10$ was found to be sufficient.



(a) Observed data with true function

(b) WPHM

(c) Our GP model

(d) The Joensuu GP model

Figure 3.2: Results from a simulated dataset with $d = 1$ generated with a squared exponential kernel with hyperparameters set to $(\eta, \beta, \sigma, l) = (5, 0.2, 3, 0.7)$. There are $N = 25$ individuals, 13 of which are censored. The end of trail at 6 years is represented by the dashed line. Figure (a) shows the observed data with the 'true' function. Figure (b) is a plot of the predicted event time using the WPHM. The grey region represents plus and minus one standard deviation from the mean prediction. We found $(\beta, \rho, \nu) = (0.49, 6.0, 4.7)$. In (c) the mean prediction using our model is shown. Optimal hyperparameters were found to be $(\eta, \beta, \sigma, l) = (5.82, 0.32, 2.59, 0.64)$. Note the increased uncertainty at $x \in (-3, -2)$. In (d) are results from the Joensuu GP model. The inferred hyperparameters are $(\eta, \beta, \sigma, l) = (-45.5, 27.2, 64.7, 0.47)$ although these are difficult to interpret since the underlying function appears as $\exp(f(x))$ in the hazard rate.

Simulated data with a specified event time density

Generation of simulated data is straightforward in the case of the GP regression model. Covariate vectors are randomly generated from a uniform distribution on a finite region of the

covariate space. The corresponding kernel matrix is constructed and event times are sampled from the GP prior which in practice means drawing a random vector from an N-dimensional multivariate Gaussian density. Censoring is simulated as above.

3.5.2 Non-monotonic simulated data example

Shown in Figure 3.2 are results from a simulated dataset that consists of $N = 25$ individuals with a single covariate x. There are 13 censored individuals and 12 who have experienced the primary risk. An end of trail cutoff at 6 years has been imposed and several individuals have been censored due to this (see Figure 3.2 (a)).

In Figure 3.2 (b) we have plotted the predicted mean event time using the WPHM. The WPHM is poorly suited to these data as it assumes a monotonically increasing or decreasing relationship between event times and covariates. The results from our model are shown in Figure 3.2 (c). The model infers the underlying function and retrieves the hyperparameters reasonably well. The inferred function gives an estimate of when event times will occur. Note that the model has extrapolated the underlying function beyond the end of trial cutoff. This can be seen in the region $x \in (-3, -2)$ and the uncertainty is also greatest in this region. The Joensuu model is also capable of inferring the underlying function quite well. Note that the uncertainty is underestimated as discussed in Section 3.4.3.

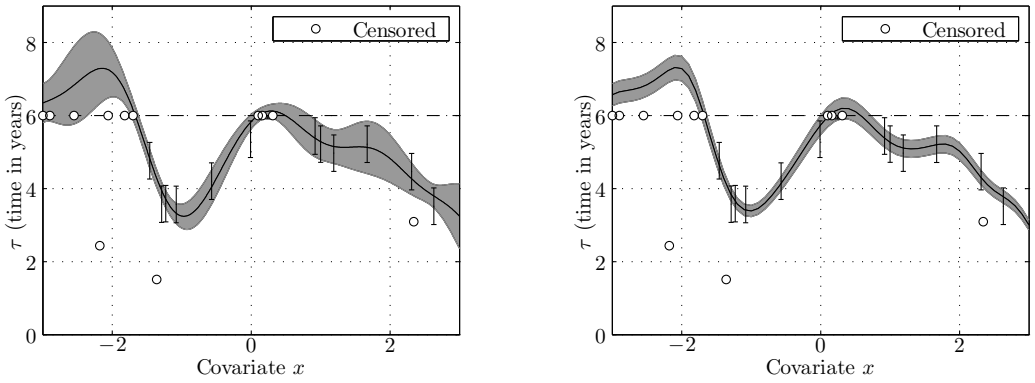
In Figure 3.3 we convert these data into interval censored data by generating a random one year interval for all of the non-censored individuals. These intervals are represented by the 'error bars' in the plot. Both GP models are capable of recovering the underlying function.

3.5.3 Monotonic simulated data example

Here we generated simulated data corresponding to the WPHM by using (3.62). These data are shown in Figure 3.4 and have a monotonic relationship between the event time and the covariate. We ran both the WPHM and our GP model in order to see how our model performs on data that can readily be analysed with existing tools. In Figure 3.4 (a) are the results from running the WPHM. Visually, it is clear that the model achieves a good fit. In Figure 3.4 (b) are the results from our GP model. Our model has also achieved a good fit. One difference between both models is that the GP model has greater uncertainty towards the left of the figure. This is appropriate since there are very few observations here so consequently our knowledge of the underlying function is less firm.

In Figure 3.4 (c) we have compared the survival functions of the WPHM, our GP model,

and a Cox proportional hazards model. The survival functions all correspond to an individual with $x = 2$. It is clear that all three models are giving broadly similar survival probabilities. Finally, in Figure 3.4 (d) we plot the hazard rates corresponding to an individual with $x = 2$ for both the WPHM and our GP model. As pointed out in Section 3.3.7 the hazard rate in the GP model is approximately linear for large times.



(a) Our GP model (b) Joensuu GP model

Figure 3.3: An example of interval censored and right censored data generated from the same function in Figure 3.2 (a). The 'error bars' denote a randomly generated one year interval. In Figure (a) are results from running our GP model. Inferred hyperparameters are $(\eta, \beta, \sigma, l) = (5.67, 0.14, 3.34, 0.57)$. In Figure (b) are results from the Joensuu GP model. Note that the uncertainty is underestimated in the Joensuu model.

3.5.4 Experimental gene expression data

We applied our method to the gene expression data from the Rosenwald et al. (2002) study of lymphoma patients. These data consist of $N = 240$ patients each with $d = 7399$ gene expression measurements. In the original analysis the patients had been split into a training group of 160 and a validation group of 80 individuals. Lu and Li (2008) studied these data to test a transformation model with non-linear covariate effects and reported that some of the gene expression levels had a non-linear relationship with the time-to-event. We examined one of these genes, with UNIQID = 33014, with our GP method and also found a non-linear function $f(\mathbf{x})$. This function was inferred using the 160 training individuals and can be seen in Figure 3.5 (a). If we compare this to the top right panel in Figure 2 of Lu and Li (2008) we can see that both functions are very similar (once we ignore the fact that, by definition,

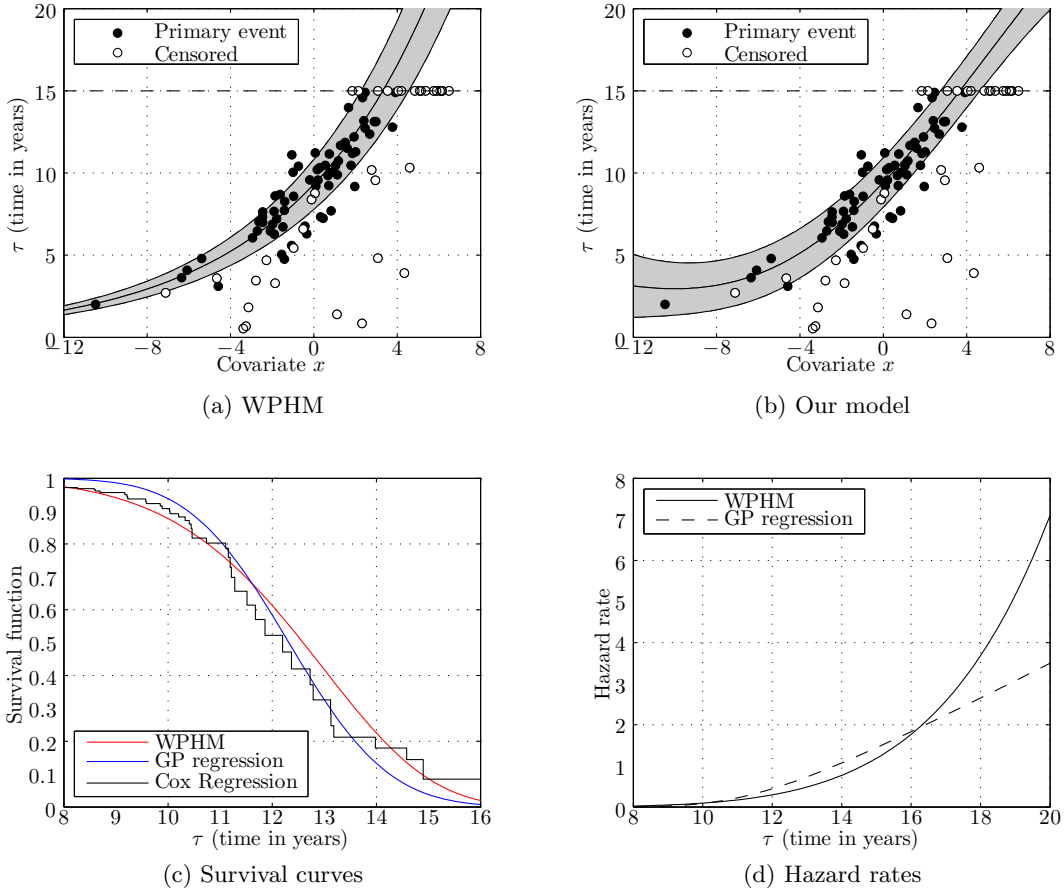


Figure 3.4: Example of data generated according to the WPHM assumptions. In (a) are results from fitting the WPHM. We found $(\beta, \rho, \nu) = (-1.04, 9.93, 7.23)$. In (b) are results from our GP model which is also capable of handling these monotonic data although the uncertainty is greater towards the left of the figure. In (c) we compare survival curves corresponding to an individual with $x = 2$ from the WPHM, our GP model and a Cox proportional hazards model (with $\beta_{cox} = -1.02$). In (d) are hazard rates from the WPHM and our GP model for an individual with $x = 2$.

they differ in sign).

To further quantify the difference between our GP method and the WPHM we computed the MSE between the predicted time-to-event and the reported time-to-event in both the training and validation sets for both models. The results are displayed in Table 4.2. It is clear that the GP method offers vastly superior performance. We can also see that the WPHM validation error is considerably larger than the training error. This is a hallmark of overfitting

where the model fails to generalise well to unseen data. GP regression on the other hand does not suffer from this problem on this dataset.

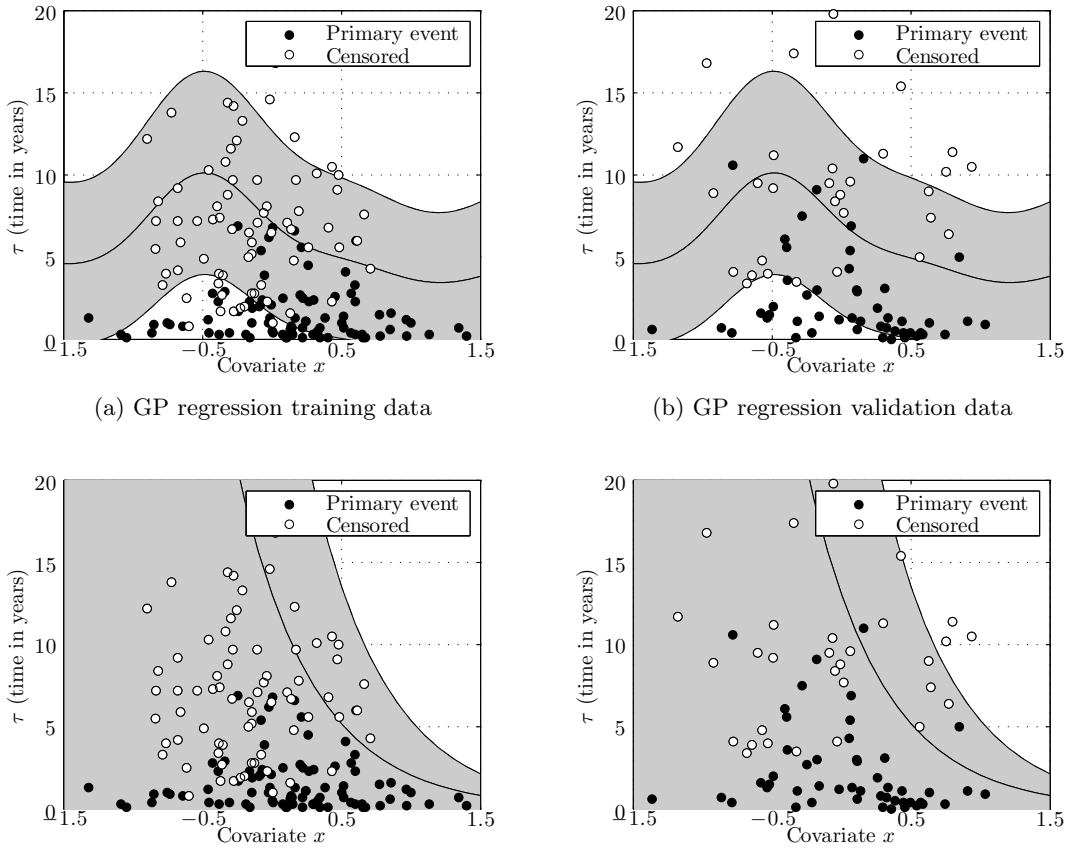
	GP regression	WPHM
Training MSE (years^2)	22.86	774.96
Validation MSE (years^2)	22.38	2514.7

Table 3.1: Comparison of mean square error between the predicted and reported event times in the validation set using gene number 33014 from the Rosenwald lymphoma dataset. Our GP regression method offers superior performance to the WPHM because it has detected a non-linear relationship between the event times and that gene expression level. The WPHM also overfits the validation data since the MSE is considerably larger than the training MSE. The GP model does not suffer from this problem in this case.

3.6 Discussion

In the case of a single risk with independent censoring the event time density, the hazard rate and the survival function are equivalent in the sense that each one of these quantities can be uniquely expressed in terms of the others. When developing methods to analyse survival data it is natural to focus on one of these quantities. Many existing methods take some parametric or semi-parametric hazard rate as their starting point. We have taken an alternative route that focuses on the event time density. Using GP regression we have formulated a highly flexible probabilistic method of relating the covariates to event times. We believe that this has a number of advantages. Firstly, it is the most direct way of connecting the two quantities we observe. It requires minimal assumptions and avoids any structural constraints that a particular hazard rate may impose. Although the GP model of Joensuu is capable of comparable performance analysing non-linear data the interpretation of hyperparameters in their model is far from obvious. This is due to the fact that they are inferring a function describing the relationship between covariates and the hazard rate. In contrast the hyperparameters in our GP model have a very clear and natural interpretation since the function that we are inferring is conceptually more straightforward and transparent.

It is relatively straightforward to include any type of censored and truncated observations (and combinations thereof). Event times can easily be estimated for censored individuals (estimates are simply the inferred function values). Standard quantities like the survival function and hazard rate can readily be obtained.



(a) GP regression training data

(b) GP regression validation data

(c) WPHM training data

(d) WPHM validation data

Figure 3.5: Univariate analysis of gene number 33014 from the Rosenwald lymphoma dataset. In (a) is the function inferred on the training set (of 160 patients) using our GP regression method which clearly shows a non-linear relationship between the expression levels and event times. In (b) is the same function superimposed on the validation set. In (c) is the inferred function on the training set using the WPHM. The WPHM clearly provides a poor fit for these data. In (d) the inferred WPHM function is superimposed on the validation set.

An interesting question to consider is how precisely to interpret the underlying function we infer. In standard GP regression the function values would be considered 'noise-free' outputs which are then corrupted by Gaussian observational noise. The corresponding interpretation in our case would be that the functions represent a 'noise free' event time. However, Gaussian noise is not appropriate in that case since it is generally not plausible to claim events could be randomly reported before they actually occur. A more appropriate choice would be noise with a semi infinite support on $(0, \infty]$ that would represent a delay between the event occurring and the time it is diagnosed or recorded. In that case the underlying function could be interpreted as a noise free event time. It may be interesting to explore different noise models in future work. Given this interpretational difficulty we may simply regard our model as a convenient way to infer a relationship between covariates and event times. Another potential research direction would be to apply sparse GP regression techniques to our model in order to achieve greater computational efficiency.

Chapter 4

Gaussian process regression with competing risks

4.1 Introduction

In this chapter we will extend the GP model developed in the previous chapter to the situation where we have competing risks. We will use multiple output GP priors which can handle the case where there are multiple potentially correlated outputs for each input. As in the case of a single risk the multiple outputs correspond to the event times for each risks.

Multiple output GP regression provides a flexible non-parametric way of modelling this. Various hyperparameters can be inferred which control the extent to which the risks are dependent. Boyle and Frean (2005) were one of the first to study the performance of multiple output GP regression. As an example they generated data with two correlated outputs. One region of the covariate space had several observations of output one but none of output two. The GP model detected the dependency between outputs and consequently was able to make much more accurate predictions for output two in that region of the covariate space than would have otherwise been possible.

The difference in the case of survival data (with two risks say) is that for each individual we will have access to at most one output per individual. This is in contrast to the more general situations multiple output GP regression could be applied to where both outputs may be observed. We also have knowledge that the other output must be greater than the observed output. This is because we know that all remaining events must have occurred sometime after the reported event time. In other words, censored observations still provide useful information.

Despite these differences our results indicate that the GP model can perform well and that predictions can be more accurate (and have smaller error bars) when the model exploits any dependence between risks.

We will begin with an overview of current methods for analysing competing risks data. Then we will provide some background information on how multiple output GP priors are constructed and implemented in Section 4.3. We will then apply multiple output GP regression to the case of two dependent risks with independent censoring in Section 4.4. A perennial question when competing risks are present is what would happen if one or more of the risks were somehow 'disabled' or 'switched off'. This issue requires careful interpretation and we discus this in detail in Section 4.5. Finally, in Section 4.6 we present results from simulation studies.

4.2 Existing methods for survival data with competing risks

Non-parametric estimates of the marginal survival curves can be generated using the Kaplan-Meier (K-M) estimator (3.1). The K-M estimator assumes that the risks are independent since events other than the event under study are regarded as independent censoring. Consequently, if the risks are indeed independent then the K-M estimators provide a valid estimate of the marginal survival probabilities. However, when the assumption of independence is incorrect then the K-M estimators cannot be interpreted as marginal survival probabilities and can lead to misleading results if they are. Andersen et al. (2012) presents some examples illustrating how the K-M estimators are biased in the presence of competing risks. However, if we are focusing on risk r, say, and all of the other cause specific hazard rates are very small then the K-M estimators will be a valid approximation of the marginal survival function. The Nelson-Aalen estimator (3.2) on the other hand is well defined in the presence of competing risks (since cause specific hazard rates are observable) but not as easy to interpret.

A common strategy is to model the cause specific hazard rates using, for example, a proportional hazards model for each rate:

$$\pi_i^r(\tau|\lambda_{0r}, \boldsymbol{\beta}_r, \mathbf{x}_i) = \lambda_{0r}(\tau)e^{\boldsymbol{\beta}_r \cdot \mathbf{x}_i}. \tag{4.1}$$

Each risk has a different vector of regression coefficients $\boldsymbol{\beta}_r$ and base hazard rate $\lambda_{0r}(\tau)$. The models are straightforward to fit since each risk can be fitted separately by treating all other events as censoring events. This follows from (2.18) and holds for both dependent and

independent risks:

$$p(D|\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_R,\lambda_{00},\ldots,\lambda_{0R}) = \prod_{r=0}^{R}\left\{\prod_{i=1}^{N}[\lambda_{0r}(\tau_i)]^{\delta_{\Delta_i,r}}e^{-\Lambda_{0r}(\tau)e^{\boldsymbol{\beta}_r\cdot\mathbf{x}_i}}\right\}. \tag{4.2}$$

The cause specific hazard rates are useful for determining the impact a covariate might have on a particular risk but it will not always be clear how the cause specific hazard rates will impact patient survival in the presence of competing risks. This is because the overall survival function and the cumulative incidence functions depend on all of the cause specific hazard rates since

$$S_i(\tau) = e^{-\sum_{q=0}^{R}\int_0^\tau \mathrm{d}s\,\pi_i^q(s)}. \tag{4.3}$$

An alternative route is to model the cumulative incidence functions:

$$C_i^r(\tau) = \int_0^\tau \mathrm{d}s\,\pi_i^r(s)e^{-\sum_{q=0}^{R}\int_0^s \mathrm{d}s'\pi_i^q(s')}. \tag{4.4}$$

As mentioned above $C_i^r(\tau)$ will in general depend on all risks and not just risk r since the probability of r occurring depends on how likely the other risks are to occur first. Since $C_i^r(\tau)$ can be written in terms of observable quantities (the cause specific hazard rates) it can be estimated from observed data. Furthermore it is easy to interpret and is useful for making predictions with.

If a proportional hazards model such as (4.1) is assumed for the hazard rates then the effect a particular covariate has on the cumulative incidence function will in general be quite complicated. Because of this models that assume a more direct relationship between the covariates and the cumulative incidence function have been proposed. A popular approach was taken by Fine and Gray (1999) who assumed a proportional hazards model

$$C_i^r(\tau) = \Phi(\Lambda_{0r}(\tau)e^{\boldsymbol{\beta}_r\cdot\mathbf{x}_i}) \quad \text{with} \quad \Phi(s) = 1 - e^s \tag{4.5}$$

that is similar in sprit to Cox's model and allows for easier quantification and interpretation of the covariate effects on the probability of succumbing to a particular risk. The $\Lambda_{0r}(\tau)$ is an unspecified monotonically increasing function of time that describes the baseline probability of failure (that is, when $\mathbf{x}_i = \mathbf{0}$). Fine and Gray (1999) interpreted this model as corresponding

to an 'effective hazard rate'

$$\pi_i^r(\tau|\mathbf{x}) = -\frac{\mathrm{d}}{\mathrm{d}\tau}(1 - C_i^r(\tau)) = \lambda_{0r}(\tau)e^{\boldsymbol{\beta}_r \cdot \mathbf{x}_i} \tag{4.6}$$

but with an unnatural risk set with an awkward interpretation. Alternatively we can simply view (4.5) as a convenient parameterisation of the cumulative incidence function. Fine (2001) subsequently extended the model to more general transformations.

Another approach is that of *relative survival*. The idea is that the hazard rate consists of two components

$$\pi(\tau) = \pi^*(\tau) + \tilde{\pi}(\tau). \tag{4.7}$$

The first component is the expected hazard rate which captures the 'background' hazard that a general population — similar to the patients under study — are exposed to. The second component is the 'excess' hazard that describes the hazard due to death from the particular disease under study. The excess hazard is usually assumed to have a proportional hazard structure and various methods of inferring the parameters have been proposed (Hakulinen and Tenkanen, 1987; Esteve et al., 1990; Dickman et al., 2004; Perme et al., 2009). It is assumed that the background hazard can be obtained from available data sources (such as population based cancer registries) after matching for age, sex and other covariates of interest so these methods are particularly useful where such large public health registries are available. The corresponding survival function is

$$S(\tau) = S^*(\tau)\tilde{S}(\tau).$$

The *net survival*, $\tilde{S}(\tau)$, is the probability of being alive at time τ in the hypothetical world where the only risk is due to the disease. For this interpretation to be valid the risk due to disease and risk due to death from other causes must be independent (this is equivalent to 'switching off' a risk in our language which we discuss in Section 4.5). Lambert et al. (2010) extend relative survival models to the competing risk scenario by computing the cumulative incidence function

$$C(\tau) = \int_0^\tau \mathrm{d}s\, \tilde{\pi}(s)S^*(s)\tilde{S}(s)$$

which gives the probability of the disease occurring sometime before τ in the presence of competing risks.

In the case of competing risks, shared frailty models have been used to capture the de-

pendencies between event times. It is assumed that there are $l = 1, \ldots, L$ known clusters of individuals and that the event times for individuals within the same cluster are dependent. A shared frailty model assumes a frailty term, w_l for each cluster, such that all individuals in that cluster share the same frailty. For example, the clusters may correspond to different trial centres in a mutli-centre trial or different studies in a meta-analysis. The intuition behind this approach is that dependencies arise between two risks due to a latent unobserved mechanism or pathway that effects both risks. Thus, the occurrence of one risk may provide information on the second risk but without directly influencing it. It's assumed that the event times within each cluster are conditionally independent given the frailty terms. Although the method is quite similar to the frailty models for single risks, conceptually the shared frailty terms represent a common dependence between the event times that isn't captured by the covariates whereas in the case of a single risk the frailty terms represent heterogeneity across individuals that isn't captured by the covariates. In a proportional hazards model for example, the hazard rate for individual i in cluster l is

$$\pi_i^l(\tau|\boldsymbol{\beta}, \mathbf{x}_i, w_l) = w_l \lambda_0(\tau) e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}.$$

See the textbook by Hougaard (2000) for more details. *Random effects models* generalise the concept of shared frailty to capture inter-cluster heterogeneity in the covariate effects (Vaida and Xu, 2000). This is achieved by assuming $\pi_i^l(\tau|\boldsymbol{\beta}, \mathbf{x}_i, w_l) = w_l \lambda_0(\tau) \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i^l + \mathbf{b}^l \cdot \mathbf{x}_i^l)$ where $\mathbf{b}^l$ is a vector of random effect coefficients that may be different for each cluster. The covariates may be further split into fixed-effect and random-effect covariates. Some distribution is typically assumed for the random effects which are then integrated out analytically if it's possible, otherwise using numerical integration techniques. It may also be important to consider models where the frailty terms are correlated with the covariates (Di Serio, 1997).

Another approach is based on pseudo-observations (for an overview see the review by Andersen and Perme (2010) and the references therein). The idea is that we are interested in some function of the event times $g(\tau_i)$ and we have an unbiased estimator $\hat{\mu}$ for the expectation of this quantity $\mu = \mathbb{E}[g(\tau)]$. For example the Kaplan-Meier estimator (3.1) is an unbiased estimator for $S(T) = \mathbb{E}[\theta(\tau - T)]$. The pseudo-observation of individual i is $\hat{\mu}_i = N\hat{\mu} - (N-1)\hat{\mu}_{-i}$ where $\hat{\mu}_{-i}$ is the estimator based on all individuals except i. The pseudo-observation can be interpreted as the contribution that that individual makes to $\mathbb{E}[g(\tau)]$. Due to censoring $g(\tau_i)$ may not be observed for each individual but pseudo-observations are available for everyone and

are used instead of $g(\tau_i)$ for all individuals. They may be subsequently used in a generalised linear regression model with a non-linear link function G

$$G(\mathbb{E}[g(\tau)]) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x} \tag{4.8}$$

which includes the Cox model and the Fine and Gray model as special cases. The can also be used to compute residuals by subtracting the pseudo-observation from the predicted value based on the model, allowing graphical goodness-of-fit tests to be made. Scatterplots of the the transformed (via the link function G) pseudo-observations and the covariates can be used to assess the appropriateness of various model assumptions (such as linear covariate effects). A potential drawback is that the estimators need to be unbiased. For example the Kaplan-Meier estimator is biased if the censoring is not independent.

Dependent competing risks can be modelled via the multivariate survival function rather than the event time density directly. A natural way to achieve this is to model the copula of dependent event times. For example, Zheng and Klein (1995) model two dependent competing risks via the copula. This is a convenient way to capture dependency between two or more risks. The authors showed that if the copula is assumed to be known then the marginal event time density can be identified and inferred from the observed survival data. One can assume a parametric form for the copula, or assume it belongs to a certain class of copulas. The latter approach is useful for defining families of multivariate survival functions with given marginal distributions. This approach requires assuming that the risks are dependent, and a particular form for the copula, in order to avoid the identifiability problem. Copulas are also implicitly used by Clayton and Cuzick (1985). Heckman and Honoré (1989) provide conditions under which dependent competing risk models can be identified. This work was further refined by Abbring and Van den Berg (2003).

4.3 Multiple output Gaussian process priors

Multiple output Gaussian process regression deals with a situation where there is more than one output variable. For simplicity we will assume there are two outputs but what follows can straightforwardly be extended to three or more outputs. Observed data consist of N_1 pairs of inputs and noisy outputs for the first output $\{(\mathbf{x}_1^1, t_1^1), \ldots, (\mathbf{x}_{N_1}^1, t_{N_1}^1)\}$ and N_2 pairs for the second output $\{(\mathbf{x}_1^2, t_1^2), \ldots, (\mathbf{x}_{N_2}^2, t_{N_2}^2)\}$. One of the main interests in this situation is to learn if both outputs are correlated and to exploit this when it comes to making predictions.

Higdon (2002) illustrated that a Gaussian process can be constructed by convolving Gaussian white noise with a kernel function. Boyle and Frean (2005) used this approach to construct two Gaussian process outputs that may be correlated. We will follow the same approach in this work and illustrate how this can be applied to the competing risks problem.

In the context of survival data the outputs are event times. We will write the event times as

$$t_1(\mathbf{x}) = f_1(\mathbf{x}) + \xi_1 \quad \text{and} \quad t_2(\mathbf{x}) = f_2(\mathbf{x}) + \xi_2, \tag{4.9}$$

with

$$f_1(\mathbf{x}) = u_1(\mathbf{x}) + s_1(\mathbf{x}) \quad \text{and} \quad f_2(\mathbf{x}) = u_2(\mathbf{x}) + s_2(\mathbf{x}), \tag{4.10}$$

where $u_r(\mathbf{x})$ is a Gaussian process unique to source r and is generated by convolving a Gaussian white noise process $l_r(\mathbf{x})$ with a kernel[1] function h_r:

$$u_r(\mathbf{x}) = \int \mathrm{d}\mathbf{z}\, h_r(\mathbf{z} - \mathbf{x}) l_r(\mathbf{z}). \tag{4.11}$$

The second component $s_r(\mathbf{x})$ is a Gaussian process generated from convolving a shared Gaussian white noise process (but possibly with a different kernel)

$$s_r(\mathbf{x}) = \int \mathrm{d}\mathbf{z}\, c_r(\mathbf{z} - \mathbf{x}) l_0(\mathbf{z}). \tag{4.12}$$

We assume that $\langle l_r(\mathbf{x}_i), l_q(\mathbf{x}_j)\rangle = \delta_{rq}\delta(\mathbf{x}_i - \mathbf{x}_j)$ for $r, q = (0, 1, 2)$. Finally Gaussian random noise $\xi_r \sim \mathcal{N}(0, \beta_r^2)$ is added to each output which represents observational noise. See Figure 4.1 for a schematic diagram.

We can now calculate the covariance between the noiseless outputs. Terms such as $\langle u_r(\mathbf{x}_i), s_q(\mathbf{x}_j)\rangle$ will vanish, leaving

$$\langle f_r(\mathbf{x}_i), f_q(\mathbf{x}_j)\rangle = \langle u_r(\mathbf{x}_i), u_q(\mathbf{x}_j)\rangle + \langle s_r(\mathbf{x}_i), s_q(\mathbf{x}_j)\rangle\,. \tag{4.13}$$

Following the example of Boyle and Frean (2005) we will assume Gaussian kernel functions defined by

$$h_r(\mathbf{z}) = \sigma_r e^{-\frac{1}{2}\mathbf{z}\cdot\boldsymbol{\Sigma}_r\mathbf{z}}, \quad c_1(\mathbf{z}) = \omega_1 e^{-\frac{1}{2}\mathbf{z}\cdot\boldsymbol{\Omega}_1\mathbf{z}} \quad \text{and} \quad c_2(\mathbf{z}) = \omega_2 e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})\cdot\boldsymbol{\Omega}_2(\mathbf{z}-\boldsymbol{\mu})}. \tag{4.14}$$

[1]The convolution kernel is related to, but distinct from the kernel function in the GP prior.

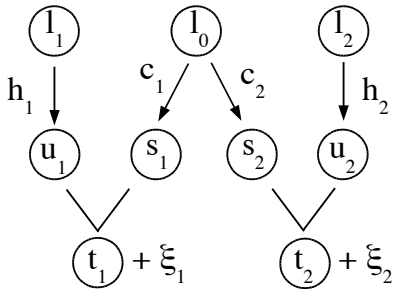


Figure 4.1: Schematic diagram of two Gaussian processes t_1 and t_2. The first output is a sum of a unique Gaussian process u_1, which is obtained by convolving the Gaussian white noise process l_1, and a shared Gaussian process s_1 which comes from convolving a shared Gaussian white noise process l_0. The second output is similarly generated using the same shared white noise process. We use h and c to denote the convolution kernels. The shared white noise process leads to dependency between outputs, which are finally corrupted by Gaussian observational noise ξ_1 and ξ_2.

The vector $\boldsymbol{\mu}$ is included to allow for an offset in both outputs. We now compute

$$\begin{aligned}\langle u_r(\mathbf{x}_i), u_q(\mathbf{x}_j)\rangle &= \int \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{z}'\, h_r(\mathbf{z}-\mathbf{x}_i)h_q(\mathbf{z}'-\mathbf{x}_j)\left\langle l_r(\mathbf{z}), l_q(\mathbf{z}')\right\rangle \\ &= \delta_{rq}\int \mathrm{d}\mathbf{z}\, h_r(\mathbf{z}-\mathbf{x}_i)h_q(\mathbf{z}-\mathbf{x}_j) \\ &= \delta_{rq}\int \mathrm{d}\mathbf{z}\, h_r(\mathbf{z})h_q(\mathbf{z}+\mathbf{d}) \end{aligned} \tag{4.15}$$

where $\mathbf{d} = \mathbf{x}_i - \mathbf{x}_j$. We can use (D.6) from Appendix D to solve these integrals and obtain

$$\langle u_r(\mathbf{x}_i), u_r(\mathbf{x}_j)\rangle = \frac{\pi^{d/2}\sigma_r^2}{\sqrt{|\boldsymbol{\Sigma}_r|}} e^{-\frac{1}{4}\mathbf{d}\cdot\boldsymbol{\Sigma}_r\mathbf{d}} \tag{4.16}$$

$$\langle s_1(\mathbf{x}_i), s_2(\mathbf{x}_j)\rangle = \frac{(2\pi)^{d/2}\omega_1\omega_2}{\sqrt{|\boldsymbol{\Omega}_1+\boldsymbol{\Omega}_2|}} e^{-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu})\cdot\boldsymbol{\Gamma}(\mathbf{d}-\boldsymbol{\mu})} \tag{4.17}$$

$$\langle s_2(\mathbf{x}_i), s_1(\mathbf{x}_j)\rangle = \frac{(2\pi)^{d/2}\omega_1\omega_2}{\sqrt{|\boldsymbol{\Omega}_1+\boldsymbol{\Omega}_2|}} e^{-\frac{1}{2}(\mathbf{d}+\boldsymbol{\mu})\cdot\boldsymbol{\Gamma}(\mathbf{d}+\boldsymbol{\mu})} \tag{4.18}$$

$$\langle s_r(\mathbf{x}_i), s_r(\mathbf{x}_j)\rangle = \frac{\pi^{d/2}\omega_r^2}{\sqrt{|\boldsymbol{\Omega}_r|}} e^{-\frac{1}{4}\mathbf{d}\cdot\boldsymbol{\Omega}_r\mathbf{d}} \tag{4.19}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Omega}_1(\boldsymbol{\Omega}_1+\boldsymbol{\Omega}_2)^{-1}\boldsymbol{\Omega}_2$. Inserting these into (4.13) allows us to construct a covariance

matrix which we can use to define a GP prior over $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2] \in \mathbb{R}^{2N}$:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}\left(\boldsymbol{\eta}, \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}\right) \tag{4.20}$$

where $[\mathbf{K}_{rq}]_{ij} = \langle f_r(\mathbf{x}_i), f_q(\mathbf{x}_j)\rangle$ and $\boldsymbol{\eta} = \eta \mathbf{I}_{2N\times 1}$, with $\eta \in \mathbb{R}$, is the GP mean. The block matrices have an intuitive interpretation. $\mathbf{K}_{11}$ and $\mathbf{K}_{22}$ control the covariance structure of the independent parts of each output whereas the off-diagonal blocks control the covariance between outputs.

Predictions

The predictive distribution for the output f_*^r corresponding to a new input $\mathbf{x}^*$ is Gaussian with mean and variance (compare to (2.51) and (2.52))

$$\hat{\mu}_r = \mathbf{k}_r^* \cdot \mathbf{K}^{-1}\mathbf{f} \tag{4.21}$$

$$\hat{\kappa}_r = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_r^* \cdot \mathbf{K}^{-1}\mathbf{k}_r^* \tag{4.22}$$

where $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2]$ and $\mathbf{k}_r^* = [\mathbf{k}_{r1}^*, \mathbf{k}_{r2}^*]$ with $[\mathbf{k}_{rq}^*]_i = \langle f^r(\mathbf{x}^*), f^q(\mathbf{x}_i)\rangle$ given by (4.13). The matrix $\mathbf{K}$ is the covariance matrix in (4.20) formed out of four block matrices. Finally, $k(\mathbf{x}_*, \mathbf{x}_*) = \langle f^r(\mathbf{x}_*), f^r(\mathbf{x}_*)\rangle = \pi^{d/2}\sigma_r^2/\sqrt{|\boldsymbol{\Sigma}_r|} + \pi^{d/2}\omega_r^2/\sqrt{|\boldsymbol{\Omega}_r|}$.

4.4 Application to two competing risks with independent censoring

We extend the transformation model from the case of a single risk (3.6) to accommodate two outputs:

$$\phi(\tau_i^1) = f_1(\mathbf{x}_i) + \xi_i^1 \quad \text{and} \quad \phi(\tau_i^2) = f_2(\mathbf{x}_i) + \xi_i^2 \quad \text{for } i = 1, \ldots, N. \tag{4.23}$$

Each event time is related to the same covariates via two different functions corrupted with two different noise random variables. In the case of competing risks the event times may be correlated so we will place a multiple output GP prior over $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2] \in \mathbb{R}^{2N}$. Throughout this thesis we will use (4.20) although alternative kernel functions could be used (Alvarez and Lawrence, 2011). Again we let $t_1 = \phi(\tau_1)$ and $t_2 = \phi(\tau_2)$ using the transformation from (3.8). The indicator variable can take values $\Delta_i = 0, 1, 2$ to indicate censoring, event type 1, and

event type 2 respectively.

A Gaussian distribution for the noise $t_r \sim \mathcal{N}(f_i, \beta_r^2)$ is assumed for $r = 1, 2$. Assuming that right censoring is independent the joint event time density is conditionally independent given the latent function values

$$p(t_i^0, t_i^1, t_i^2 | f_i^1, f_i^2) = p(t_i^0) p(t_i^1 | f_i^1) p(t_i^2 | f_i^2). \tag{4.24}$$

The conditional independence leaves us with a rather convenient event time density. All of the complicated business of correlations between risks and similarities between individuals is captured by the GP prior leaving a simple product of univariate Gaussian densities. We will discus thus more in Section 4.5. The probability of censoring $p(t_i^0)$ is assumed constant and can be absorbed into the normalisation factor and henceforth will be ignored. Due to the conditional independence we can use the formulae for independent risks (Section 2.1.3). From (2.20) the survival function is given by $S_i(t|f_i^1, f_i^2) = S_i^1(t_i|f_i^1) S_i^2(t_i|f_i^2)$ and from (2.22) the cause specific hazard rates are $\pi_i^r(t|f_i^r) = p_i^r(t|f_i^r)/S_i^r(t|f_i^r)$.

4.4.1 Interpretation of hyperparameters

We will make a number of simplifying choices for the hyperparameters. We assume $\sigma = \sigma_1 = \sigma_2$ which controls the variance of the outputs that is due to their unique Gaussian processes. Secondly, the hyperparameters $\omega = \omega_1 = \omega_2$ control the variance due to shared latent processes. These parameters control to what extent the outputs are dependent. In the case where $\omega = 0$ the sub matrices $\mathbf{K}_{12}$ and $\mathbf{K}_{21}$ in (4.20) are zero. Consequently, $\mathbf{f}^1$ and $\mathbf{f}^2$ are independent and it follows that the prior density factorises over risks. In this case the two risks are completely independent and it is equivalent to fitting two separate functions for each risk independently (albeit with the same hyperparameter σ). In the results section we will sometimes fix $\omega = 0$ in order to compare a model that assumes independent risks to a model that allows for dependencies to exist.

Together σ and ω control the overall variance of $\mathbf{f}$, that is, the overall timescale over which we expect to see events occurring. It is not unreasonable to assume that these timescales will be similar in both risks as otherwise there is little motivation for a competing risks analysis as one event type will preclude the observation of the second type.

We assume that the covariance matrices in (4.14) are diagonal with $\mathbf{\Sigma}_r = \Sigma_r \mathbf{I}_{d\times d}$ and $\mathbf{\Omega}_r = \Omega_r \mathbf{I}_{d\times d}$. We assume that the characteristic length scales in covariate space are the same $\Sigma_1 = \Sigma_2 = \Omega_1 = \Omega_2 = 1/l^2$. Finally, we shall assume that the noise levels are the

same for both events $\beta_1 = \beta_2 = \beta$. In the simplest case we have a six-dimensional vector of hyperparameters $\boldsymbol{\theta} = (\eta, \mu, \beta, \sigma, \omega, l)$ where $\eta \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\beta \geq 0$, $\sigma \geq 0$, $\omega \geq 0$, and $l \geq 0$. These simplifications are by no means necessary and may not be appropriate for certain datasets. They do however make inference of hyperparameters considerably easier since the search space will in general contain many local minima so making the dimension of that space as small as possible will have significant computational advantages.

4.4.2 Inference of latent function values and hyperparameters

Due to the convenient conditional independence of the event times (4.24) we can work directly with (2.25) except that in this case we ignore any terms corresponding to right censoring since this is assumed to be independent and random:

$$p(D|\mathbf{f}_1, \mathbf{f}_2) = \prod_{r=1}^{2} \left\{ \prod_{i=1}^{N} [p_i^r(\tau_i)]^{\delta_{\Delta_i, r}} [S_i^r(\tau_i)]^{1-\delta_{\Delta_i, r}} \right\}. \tag{4.25}$$

As before, we use Bayes' theorem to calculate the posterior over the latent function values:

$$p(\mathbf{f}|\mathbf{X}, D) = \frac{p(D|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{p(D|\mathbf{X})}. \tag{4.26}$$

where $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2]$. We define the negative log likelihood as

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N} \log p(\mathbf{f}|\mathbf{X}, D).$$

We can then substitute terms from (4.25) and (4.20) into the negative log likelihood to obtain

$$\begin{aligned} \mathcal{L}(\mathbf{f}) = &-\frac{1}{N} \sum_{i:\Delta_i \neq 1} \log S_i^1(t_i|f_i^1) - \frac{1}{N} \sum_{i:\Delta_i \neq 2} \log S_i^2(t_i|f_i^2) - \frac{1}{N} \sum_{i:\Delta_i = 1} \log p_i(t_i|f_i^1) \\ &- \frac{1}{N} \sum_{i:\Delta_i = 2} \log p_i(t_i|f_i^2) + \frac{1}{2N}(\mathbf{f} - \boldsymbol{\eta}) \cdot \mathbf{K}^{-1}(\mathbf{f} - \boldsymbol{\eta}) + \log 2\pi + \frac{1}{2N} \log |\mathbf{K}|. \end{aligned} \tag{4.27}$$

The Laplace approximation of the marginal posterior density is constructed and used to define the negative log hyperparameter posterior

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) + \frac{1}{2N} \log |N\mathbf{H}| \tag{4.28}$$

where $\hat{\mathbf{f}} = \min_{\mathbf{f}} \mathcal{L}(\mathbf{f})$. First and second order partial derivatives are given in Appendix A.5. Note that it is straightforward to show that the posterior (4.26) is proper by following a similar argument to Section 3.3.4.

4.4.3 Making predictions

The predictive distribution corresponding to an individual with covariates $\mathbf{x}^*$ is $p(t_*^r|\mathbf{x}^*, \mathbf{X}, D) = \mathcal{N}(\hat{\mu}_r, \hat{\kappa}_r + \beta_r^2)$ with $\hat{\mu}_r$ and $\hat{\kappa}_r$ given by (4.21) and (4.22) respectively. The predictive density over the original event time variable is

$$p(\tau^*|\mathbf{x}^*, \mathbf{X}, D) = \frac{e^{-\frac{1}{2(\hat{\kappa}+\beta^2)}(\log(e^{\tau^*/\gamma}-1)-\hat{\mu})^2}}{(2\pi(\hat{\kappa}+\beta^2))^{1/2}} \frac{e^{\tau^*/\gamma}}{\gamma(e^{\tau^*/\gamma}-1)}. \tag{4.29}$$

From this the mean and variance can be numerically computed, similarly to Section 3.3.5. Once the predictive event time density has been obtained one can readily derive hazard rates or survival curves if desired.

4.5 'Disabling' a risk

A question of great interest in a competing risks situation is how to estimate the survival probabilities from one risk in the absence of one or all of the other risks. It has been described as *the problem* of competing risk by Kalbfleisch and Prentice (2002, Chapter 8) and has at least been asked since Bernoulli (1760). It is more than just a statistical question since 'switching off' or disabling some of the risks will in general alter the hazard rates of the remaining risks because the risks will in general share common biological pathways or depend to some extent on the same components. To interfere with the underlying system so as to disable some of the risks will in general change the characteristics of the remaining risks. It may be very difficult or impossible to quantify these changes in practice. In any case, the data we have available comes from a system where all risks are operating and therefore it is in general not possible to predict what effect switching off one of more of the risk will have.

We can however imagine a hypothetical world where some of the risks have been disabled and the hazard rates of the remaining risks are still relevant. To be clear, the assumption is that hazard rates inferred from a situation in which all risks are operating are applicable to a situation where one or more of those risks have been disabled. Regardless of whether or not this hypothetical situation is plausible it is nonetheless a case than can be addressed

statistically. The precise interpretation of marginal survival probabilities and cumulative incidence functions depends on whether the risks are independent or not.

We now provide the mathematical details. By 'switching off' all risks except risk r we mean replacing

$$p_i(\tau_0, \ldots, \tau_R) = \tilde{p}_i^r(\tau_r) \lim_{\zeta \to \infty} \prod_{q \neq r} \delta(\tau_q - \zeta) \tag{4.30}$$

where $\tilde{p}_i^r(\tau) = \int_0^\infty (\prod_{q \neq r} \mathrm{d}s_q) p(s_0, \ldots, s_R)$ is the marginal density of event time r. We use tildes to denote quantities after the risks have been disabled. If we substitute (4.30) into (2.3) we get

$$\tilde{S}_i^r(\tau) = \int_{\tau_r}^\infty \mathrm{d}s\, \tilde{p}_i^r(s). \tag{4.31}$$

The cause specific hazard rate can be obtained from (2.27) since there is only one risk:

$$\tilde{\pi}_i^r(\tau) = \tilde{p}_i^r(\tau) / \tilde{S}_i^r(\tau). \tag{4.32}$$

From this it follows that

$$\tilde{S}_i^r(\tau) = e^{-\int_0^\tau \mathrm{d}s\, \tilde{\pi}_i^r(s)}. \tag{4.33}$$

Note that $\tilde{\pi}_i^q(\tau) = 0$ and $\tilde{S}_i^q(\tau) = 1$ for all $q \neq r$. The cumulative incidence function for risk r becomes

$$\tilde{C}_i^r(\tau) = \int_0^\tau \mathrm{d}s\, \tilde{\pi}_i^r(s) e^{-\int_0^s \mathrm{d}s' \tilde{\pi}_i^r(s')} = 1 - \tilde{S}_i^r(\tau). \tag{4.34}$$

We now examine the case of dependent and independent risks separately.

Dependent risks

In this case $\tilde{\pi}_i^r(\tau) \neq \pi_i^r(\tau)$. This can be seen by comparing the expression for $\pi_i^r(\tau)$ given by (2.4) to the expression for $\tilde{\pi}_i^r(\tau)$ which is given by (4.32). This is to be expected since switching off the other risks will change the probability to survive until a certain time and hence the hazard due to risk r will also change. In this case the quantity $\exp(-\int_0^\tau \mathrm{d}s\, \pi_i^r(s))$ cannot be interpreted as a marginal survival probability in the hypothetical world where all other risks are switched off. Consequently, $C_i^r(\tau) = 1 - S_i^r(\tau)$ does not have a valid interpretation as a cumulative probability distribution either.

Independent risks

In the case of independent risks the survival function can be written as $S_i(\tau) = S_i^1(\tau) \cdots S_i^R(\tau)$ where the marginal survival functions are defined as $S_i^r(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i^r(s)$ for $r = 1, \ldots, R$. Since the risks are independent it immediately follows that $\tilde{p}_i^r(\tau) = p_i^r(\tau)$. From (4.31) and (4.32) it follows that $\tilde{S}_i^r(\tau) = S_i^r(\tau)$ and $\tilde{\pi}_i^r(\tau) = \pi_i^r(\tau)$. In this case the quantity $S_i^r(\tau) = \exp(-\int_0^\tau \mathrm{d}s\, \pi_i^r(s))$ is equal to (4.33) and hence it can be interpreted as a marginal survival probability in the hypothetical world where all other risks are switched off.

The GP model

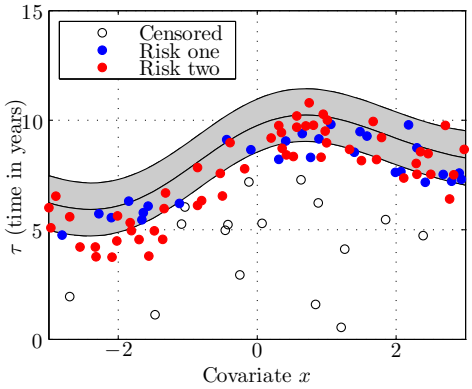
In our case the conditional independence of the event times given the latent function means that we can always interpret $S_i^r(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i^r(s|f_i^r)$ as a marginal survival probability. This is true regardless of whether the underlying functions are independent or otherwise (which in the language of our model means this is true for any value of ω).

4.6 Results

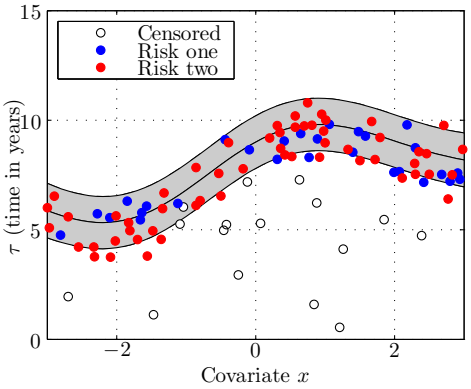
We study the behaviour of the GP model under a variety of conditions using simulated data. We also compare the performance of our GP model to that of the WPHM.

4.6.1 Non-monotonic survival with dependent competing risks

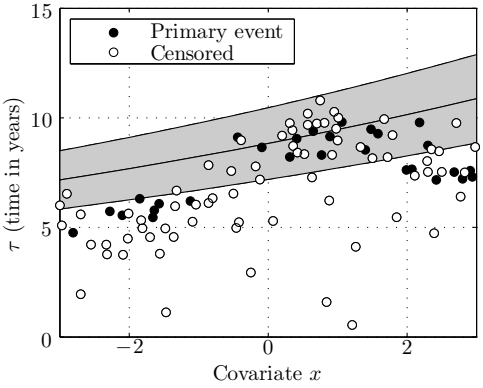
Shown in Figure 4.2 are non-monotonic simulated data with two dependent competing risks. A total of $N = 100$ samples are generated from two functions drawn from a multiple output GP prior. The GP model assumes that the risk are dependent by inferring a value of $\omega = 1.2$ for the shared variance compared to $\sigma = 0.25$ for the unique variance. In the same figure we show results from running two independent WPHMs. As expected the WPHM is unable to handle the non-monotonic nature of these data. To quantify this we generated a further 100 validation samples. Using the trained GP model and the trained WPHM models we made predictions for validation samples as follows. For individuals who reported event type 1 first we made a prediction of the time until that event. If event type 2 was reported then we predicted the time until event 2. Censored individuals were excluded. We then compute the mean square error (MSE) between predicted and reported event times. Results are shown in Table 4.1.



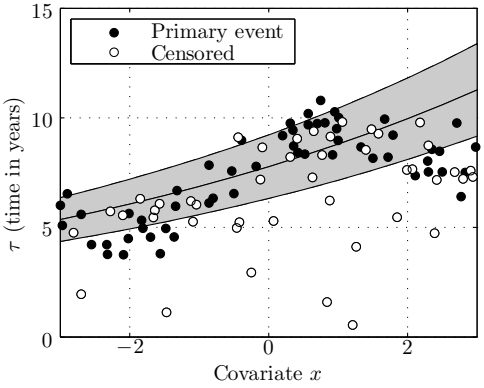
(a) GP model inferred risk 1



(b) GP model inferred risk 2



(c) WPHM inferred risk 1



(d) WPHM inferred risk 2

Figure 4.2: Example of non-monotonic survival data with two dependent competing risks. In Figures (a) and (b) are plots of the inferred risk one and risk two functions respectively using the GP model. Inferred hyperparameters are $(\eta, \mu, \beta, \sigma, \omega, l) = (7.93, -0.16, 1.18, 0.25, 1.20, 1.15)$. The value $\omega \neq 0$ reflects the fact that the model has assumed the risks are dependent. In Figures (c) and (d) are results from running two independent WPHM models. In each model only one of the risks is regarded as the primary risk and all other events are considered as right censored.

	WPHM	GP regression
Risk 1 MSE (years^2)	2.43	1.42
Risk 2 MSE (years^2)	2.67	0.70

Table 4.1: Comparison of mean square error between the WPHM and the GP model on 100 validation samples corresponding to the training data in Figure 4.2. The GP model has considerably better predictive performance.

4.6.2 Monotonic survival data with dependent competing risks

We then repeated the entire experiment but with data that follow a roughly monotonic pattern. These data are shown in Figure 4.3. Again, we train a GP model and compare this to two independent WPHM models. In this case values of $\omega = 0.95$ and $\sigma = 0$ were inferred which indicates that the risks are completely dependent with no variance due to unique components. The characteristic length scale $l = 4.56$ illustrates that the function is changing relatively slowly with respect to the covariate. Compare this to a value of $l = 1.15$ from the example in Figure 4.2. In Figure 4.3 (a) we see that the GP model has inferred a risk 1 function that lies 'above' most of the observed event times. This is not unreasonable since all of the risk 2 events are effectively censoring events from the point of view of risk 1. Therefore, we know that risk 1 events must occur after risk 2 event times. Consequently the data likelihood is maximised by placing the risk 1 function slightly 'above' the risk 2 events. A similar effect can be seen in Figure 4.3 (c) since the risk 2 events are also regarded as censoring events in the WPHM. As in the example above we compute the MSE in a validation set of 100 samples. Results are shown in Table 4.2. The GP model performs slightly worse that the WPHM model, particularly when it comes to predicting risk 1 events. This appears to be due to the fact that the GP model has inferred a risk 1 function that is slightly 'higher' than the WPHM risk 1 function. Consequently, the predicted risk 1 events are overestimating the time to event. In this dataset there are 66 risk 2 events compared to 19 risk 1 events which helps to explain the poorer predictions for risk 1. Also, because of the way the validation data are generated only the earliest risk 1 events are reported which leads to larger MSE values in both models because the predicted event times tend to be later than the reported event times.

	WPHM	GP regression
Risk 1 MSE (years^2)	1.42	3.10
Risk 2 MSE (years^2)	0.75	0.84

Table 4.2: Comparison of mean square error between the WPHM and the GP model on 100 validation samples corresponding to the training data in Figure 4.3. The GP model has slightly poorer performance than the WPHM, particularly on risk 1. See the main text for further discussion.

4.6.3 Comparison of GP models with dependent and independent risks

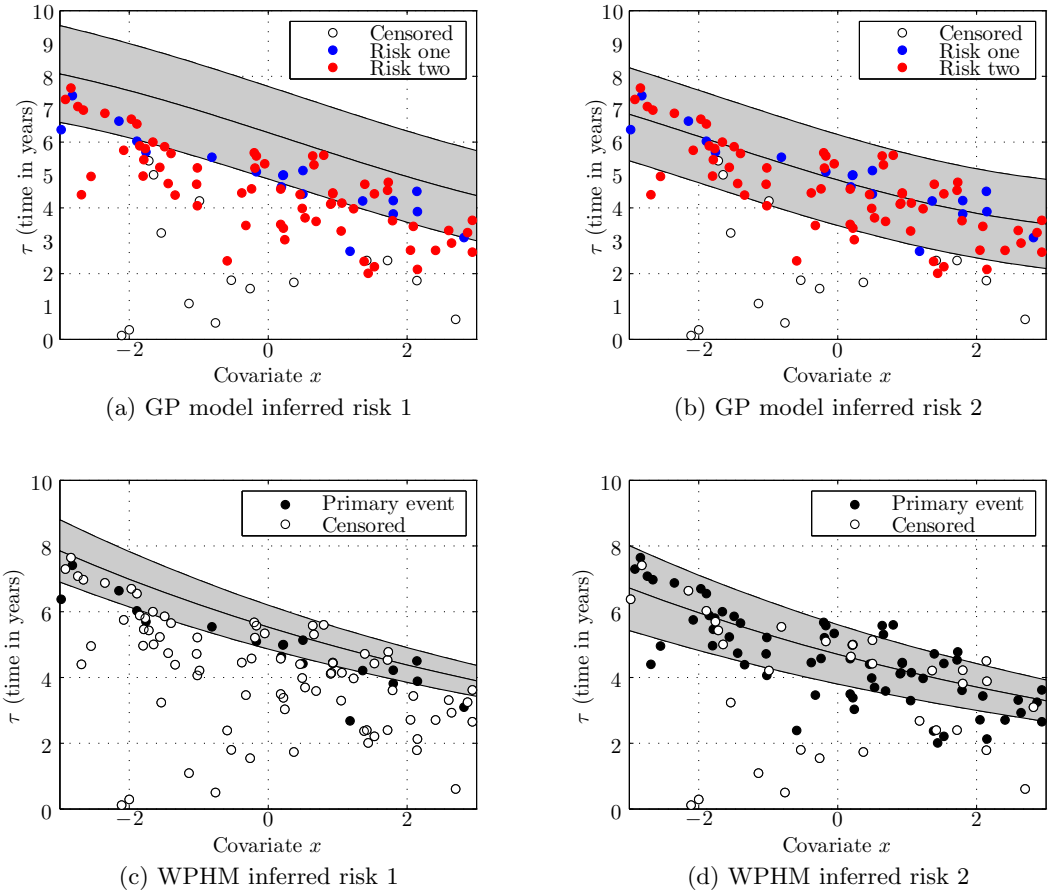
By fixing the value of $\omega = 0$ we force the two risks to be independent in the GP model. We generated survival data with dependent risks and compare two GP models, one which allows for dependency between risks and one with independent risks. The results are shown in Figure 4.4.

In (a) and (b) are results from a GP model where ω is inferred from the data. $(\eta, \mu, \beta, \sigma, \omega, l) = (4.59, 0.41, 0.33, 0.20, 1.63, 1.01)$. The higher value of ω indicates that the model is assuming strong dependence between risks. In (c) and (d) are results from a second GP model with $\omega = 0$. Remaining hyperparameters were found to be $(\eta, \mu, \beta, \sigma, l) = (13.03, -0.60, 1.02, 1.90, 1.02)$. Note that the value of σ is now higher as the unique part of each risk must explain all of the 'output' variance. The advantage of allowing dependent risks becomes apparent when we examine the inferred risk 2 function towards the left of (b) and (d). In the independent model the uncertainty associated with the underlying function is much greater since knowledge of risk 1 is unavailable. In the dependent model a more accurate recovery of the risk 2 function is obtained and the uncertainty is smaller since information from risk 1 events can be utilised more effectively.

In Figure 4.4 (e) we compare the two inferred risk 2 functions to the 'true' risk 2 function. We can see that the dependent model has done quite a good job at recovering the correct function despite the complete lack of risk 2 observations to the left of the x-axis. Of course, with real data we will not have the luxury of knowing whether an assumption of dependence is correct or not but this example at least illustrates the potential usefulness of our approach.

4.6.4 Example of two dimensional covariates

The model is fully capable of dealing with multi-dimensional covariates. By using the squared exponential kernel with automatic relevance determination (ARD) hyperparameters (see Sec-



(a) GP model inferred risk 1

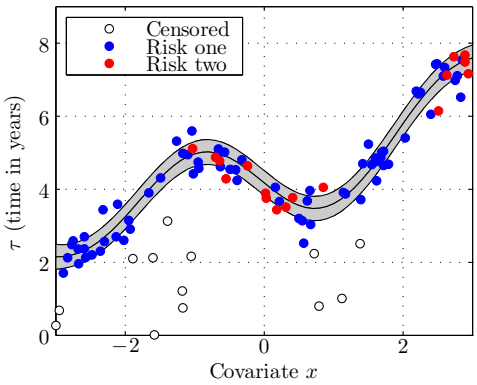
(b) GP model inferred risk 2

(c) WPHM inferred risk 1

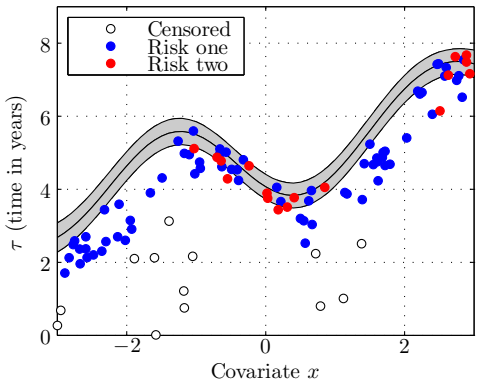
(d) WPHM inferred risk 2

Figure 4.3: Example of monotonic survival data with two dependent competing risks. In Figures (a) and (b) are plots of the inferred risk one and risk two functions respectively using the GP model. Inferred hyperparameters are $(\eta, \mu, \beta, \sigma, \omega, l) = (5.95, 2.16, 1.40, 0, 0.96, 4.56)$. The value $\omega \neq 0$ reflects the fact that the model has assumed the risks are dependent. The characteristic length scale $l = 4.56$ indicates that the function is changing relatively slowly with respect to the covariate. In Figures (c) and (d) are results from running two independent WPHM models. In each model only one of the risks is regarded as the primary risk and all other events are considered as right censored.

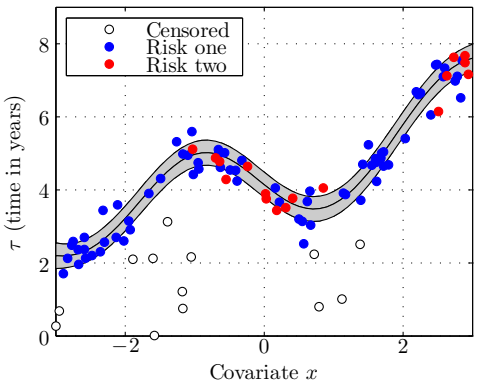
tion 2.3.1) we can determine which covariates are the most important. This is analogous to examining the regression coefficients in a Cox model to see which covariates have the greatest impact on survival outcomes. In the example shown in Figure 4.5 we find that $(l_1, l_2) = (0.52, 1.47)$ indicating that the first covariate is more important.



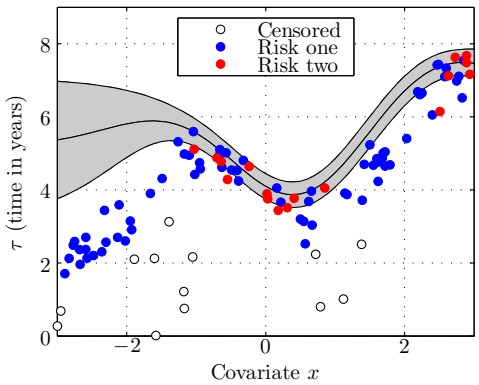
(a) Dependent risk 1



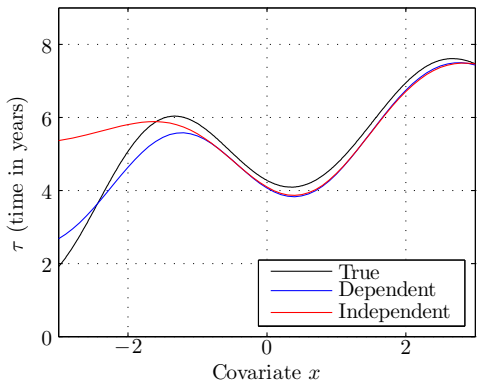
(b) Dependent risk 2



(c) Independent risk 1

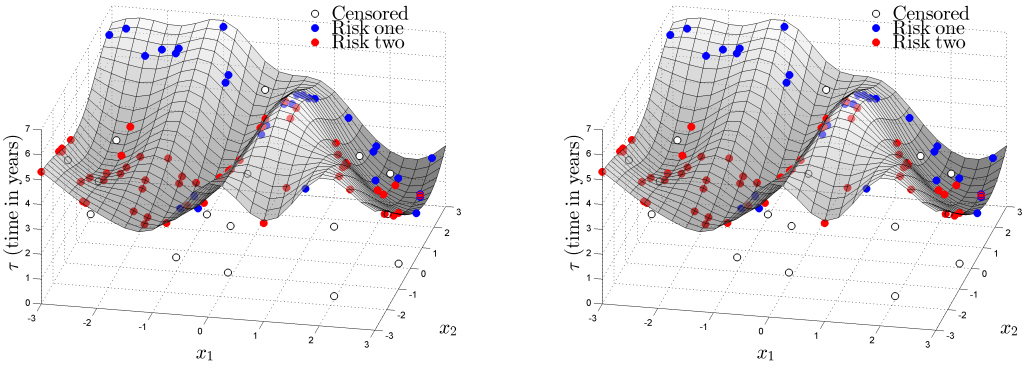


(d) Independent risk 2



(e) Comparison

Figure 4.4: In (a) and (b) are results from a GP model with dependent risks allowed. In (c) and (d) the inferred risks are forced to be independent by setting $\omega = 0$. In (e) is a comparison of both inferred functions for risk 2 compared to the 'true' function. See the main text in Section 4.6.3 for full details.



(a) Inferred risk 1

(b) Inferred risk 2

Figure 4.5: Example of GP regression with two dimensional covariates and strongly dependent risks. Covariate x_1 has an inferred characteristic length scale of $l_1 = 0.52$ compared to $l_2 = 1.47$. The values used to generate the data were $(l_1, l_2) = (0.5, 1.5)$. This indicates that the first covariate is more relevant to determining survival outcomes. This is reflected in the figures since the function is more variable in the x_1 direction. The remaining hyperparameters were found to be $(\eta, \beta, \sigma, \omega) = (4.83, 0.17, 0.23, 0.89)$. Note that higher value of ω reflects the fact that the GP model has assumed a strong dependence between risks.

4.7 Discussion — why is survival analysis hazard based?

Modelling the hazard rate has arguably provided the most popular route to analysing survival data. Cox's proportional hazards model and its myriad variations are classic examples of this approach. To a lesser extent the cumulative incidence function has also proved to be a popular approach. The appeal of these approaches is easy to understand. The hazard rates and cumulative incidence functions can be inferred from observed data, have an intuitive interpretation and are useful for establishing associations between covariates and survival outcomes and making predictions about future events.

The latent failure time approach, where one thinks in terms of the joint probability density of the failure times, is an alternative approach to analysing survival data. This approach has been criticised for a number of reasons. The most serious objection is that the joint event time density is unobservable and cannot be inferred from the observed failure and censoring times. While one may assume that the event times are independent or perhaps assume some parametric density to model dependencies one cannot conclude on the basis of the observed data alone whether or not the event times are independent. Some authors such as Beyersmann

et al. (2012, Section 3.3) have claimed that the latent failure times lack plausibility or are too hypothetical in nature since we are positing the existence of quantities which in reality can never be measured (see Kalbfleisch and Prentice (2002, Section 8.2) for further in depth discussion).

From this point of view the hazard rates provide a more natural and intuitive framework for analysing survival data. In the real world we observe events that necessarily occur to individuals that are still alive and the hazard rate captures precisely this — the rate at which events occur to living individuals. Being wholly observable quantities the hazard rates can be inferred from observed data and do not require any assumptions to be made regarding the independence of the risks. In addition, Crowder (2012, Section 14.3) points out that it is much easier to incorporate time dependent covariates using hazard rates than via the joint event time density. For similar reasons the cumulative incidence function has been a popular alternative to a hazard rate formulation of survival analysis. Beyersmann et al. (2012, Section 3.3) conclude that:

> "Assuming latent times rather appears to cause confusion and to create artificial problems than to contribute to an understanding of the subject matter. Although logically feasible, the concept of latent times is not convincing, except for special cases such as a technical device whose single components have a physical and functional interpretation. Except for such special cases, the latent failure time model is philosophically rather 'expensive', assuming, e.g., that a human being is equipped with a large enough reservoir of latent times for any competing risks situation that the individual might face."

We argue instead that the latent failure time approach does provide a useful conceptual framework. There are two aspects in particular where such an approach is useful. The first is in making predictions for new patients and the second is in estimating what happens when one or more risks are disabled. In both cases the marginal survival functions are the relevant quantities. Firstly, while it may be implausible to consider the latent failure times for an individual who has already died from one of the events the situation is rather different for individuals who are still alive. The time until different events is highly relevant for making predictions and clinical decisions. As soon as we ask questions about several risks it is natural to consider the event times as random variables and it follows that we are immediately interested in the joint probability density. As we have shown here our GP model provides a straightforward way to predict the time-to-event for new individuals. If we want

to model dependent risks (despite not being able to test our assumption) then modelling the joint event time density is a convenient starting point. The fact that the data we observe in reality do not allow a direct view of this joint density does not mean that it is not a useful concept.

Secondly, it provides a useful framework when considering what would happen when a risk is 'switched off'. The conditional independence of the latent event time density allows us to easily predict marginal survival probabilities after switching off a risk. The difficulties in understanding what it means to switch off a risk are twofold. Firstly, it requires an assumption that the hazard rates inferred under conditions where all risks are present will be relevant to conditions where one or more of those risks have been somehow eliminated. This is not primarily a statistical question since it will depend on the nature of the underlying physical or biological system. Indeed, any extrapolation to a scenario where no data have been observed will be problematic for any statistical modelling. Kalbfleisch and Prentice (2002, Section 8.2.6) summarise:

> "A statistical specification of failure rates given the removal of certain failure types will be sensible only in very special cases. In some specific applications, it may be possible to utilise existing data to reach sensible inferences. A detailed knowledge of the system giving rise to the failures and knowledge of the removal mechanism is required for such an extrapolation."

A further issue concerning the disabling of one or more risks is that the interpretation of the marginal survival probabilities will depend on whether the risks are independent or not. Of course, this cannot be inferred from the data and must be assumed. Within the framework of our GP model the marginal survival probabilities have a valid interpretation regardless of whether the model assumes the risks are independent or not.

Another advantage of our latent event time approach is that it avoids any explicit structural assumptions of what form the hazard rate takes. Modelling the hazard rate requires capturing both time and covariate effects whereas modelling the joint event time density requires capturing only the covariate effects. Consequently, specifying the event times as a function of the covariates requires fewer assumptions. Using GP regression results in a highly flexible non-parametric probabilistic model. The default model settings allows for dependent risks; therefore, an assumption of dependence is made. As our results have shown assuming dependence between risks can lead to more accurate predictions in some cases. Future directions of research are similar to the GP model with a single risk. It would be interesting to

explore different noise distributions and to apply some of the sparse GP techniques to achieve greater computational efficiency.

Chapter 5

The Gaussian process latent variable model

5.1 Introduction

High dimensional data are those where the number of covariates d is relatively large compared to the number of individuals N. The analysis of high dimensional data is challenging due to the risk of *overfitting*. Overfitting occurs when we infer spurious relationships between data that are not biologically genuine and have occurred by chance. In order to measure overfitting in situations where we wish to relate covariates to some output variable (time-to-event or class membership labels for example), we typically split our dataset into a training set and a validation (or test) set. A model is fitted using the training data, usually by inferring the values of some model parameters, and is then tested by using the trained model to predict the outcomes associated with each individual in the validation set. A hallmark of overfitting is to observe a good fit to the training dataset, but poor performance on the validation set since the model has picked up on spurious relationships that fail to exist in the validation data. The greater the dimension of the covariates compared to the number of individuals the greater the risk of picking up spurious patterns, as the model struggles to find meaningful patterns in such large volumes of data.

High dimensional data are becoming increasingly common in biomedical research. New experimental techniques can easily generate a huge number of measurements for a single individual. Gene expression microarrays can acquire expression levels for tens of thousands of genes in a single experiment. Single nucleotide polymorphism (SNP) data or DNA methylation

data consist of hundreds of thousands of measurements. Automated image analysis software can extract hundreds or thousands of parameters from various types of biomedical imaging platforms. For example, parameters that quantify the texture, shape, and density of tumours can be extracted from PET scans. Fluorescent microscopy techniques can be used to see the presence of different proteins in thin slices of cancer tumours. Förster resonance energy transfer (FRET) is a mechanism that allows researchers to infer the distance between two proteins which indicates whether two proteins are interacting. Parameters that characterise the abundance, spatial distribution and pairwise interaction of proteins can be extracted from these fluorescent images. All of these data types have high dimensional covariates in common. Typical experimental datasets may have in the order of ten or one hundred patients and in rare cases thousands which means that the danger of overfitting is frequently present. The so-called 'curse of dimensionality' has become a common problem.

Dimensionality reduction methods attempt to address the overfitting problem by representing the information in a dataset in a lower dimensional space. By lowering the ratio of covariates to samples we diminish the risk of overfitting. There are numerous other strategies for dealing with high dimensional data and in the following chapter we will discus these in more detail. For the moment we focus on one particular dimensionality reduction model called the Gaussian process latent variable model (GPLVM).

The GPLVM was originally introduced by Lawrence (2005) as a flexible probabilistic dimensionality reduction method. For individual i we observe a high dimension vector of covariates $\mathbf{y}_i \in \mathbb{R}^d$ which we attempt to represent with a vector of latent variables $\mathbf{x}_i \in \mathbb{R}^q$. There are N individuals in total. Ideally the number of latent variables q will be much smaller than the dimension of the observed data d, thus achieving a more parsimonious representation of the data. The GPLVM is closely related to probabilistic principal component analysis (Tipping and Bishop, 1999) which itself is a generalisation of factor analysis (see the textbook by Rencher (2002)).

The GPLVM has several attractive features. It assumes that the components of the high dimensional data $\mathbf{y}_i$ can be written as functions of $\mathbf{x}_i$, and Gaussian process (GP) priors are assumed for these functions. By choosing different kernel functions in the GP prior we can specify different types of non-linear mappings between the low and high dimensional spaces. This allows for a highly flexible model that can generate non-linear embeddings in the latent space.

In the GPLVM the latent variable representations $\mathbf{x}_i$ are unknown and must be inferred from the data. In general this must be done numerically. However the simplest case, where

a linear kernel is used, corresponds to a linear mapping from the latent variables to the high dimensional covariates. It was shown by Lawrence (2005) that in this case the maximum a posteriori (MAP) solution can be obtained analytically and is equivalent to performing Principal Component Analysis (PCA) and retaining the first q principal components. Intuitively we can regard the GPLVM as a non-linear generalisation of PCA when non-liner kernel functions are used.

One disadvantage of the original GPLVM is its computational complexity. This is primarily due to the need to invert an $N \times N$ kernel matrix, where N is the number of samples in the dataset. This issue was addressed by Lawrence (2007) by applying sparse GP regression methods to the GPLVM. Sparse GP regression (Snelson and Ghahramani, 2006) uses various approximations to reduce the computational complexity required during inference. A thorough overview of sparse GP regression can be found in Quiñonero-Candela and Rasmussen (2005).

Building on developments in variational sparse GP regression (Titsias, 2009), the Bayesian Gaussian Process Latent Variable Model was introduced by Titsias and Lawrence (2010). The idea behind variational approaches is to approximate the marginal distribution (obtained by integrating out the latent variables) with a lower bound. The lower bound is then made to 'fit tightly' by minimising the Kullback-Leibler divergence between the lower bound and the true distribution with respect to hyperparameters and variational parameters. Usually, the variational bound is easier to evaluate and optimise with respect to hyperparameters. A detailed description of the variational approach to the GPLVM can be found in Gal et al. (2014). This approach can also be used to detect the intrinsic dimensionality of the latent variable space. This is achieved by using a squared exponential kernel with automatic relevance determination (ARD) hyperparameters. Latent variables that are not required can be 'switched off' by sending their corresponding ARD parameters to zero.

Another natural extension of the GPLVM is to incorporate multiple datasets simultaneously. If we are to acquire two datasets, say $\mathbf{Y}_1 \in \mathbb{R}^{N \times d_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times d_2}$, it may be that there is overlap in the content and structure of these data. Suppose we were to obtain gene expression data for a cohort of patients and also parameters extracted from fluorescent imaging data. It may be the case, for example, that individuals with disrupted gene expression levels may have altered levels of protein abundance or interaction (due to the underlying biological processes). It is desirable to somehow combine the information from both of these sources; particularly if any signal in the data can be strengthened. The GPLVM can naturally accommodate these types of data by expressing both datasets in terms of the same latent variables.

The shared latent variables then capture the shared structure between multiple sources.

This idea was developed by Shon et al. (2006) and Ek et al. (2007) who called their model the shared-GPLVM. They applied the model to image synthesis where images of the same object are taken from different angles. A disadvantage of this approach is that sometimes each dataset may have some information that is unique to it. This problem was overcome by Ek et al. (2008) by dividing the space of latent variables into shared and private parts. The private subspaces are available only for each dataset and describe variability that is unique to that particular source while the shared latent variables extract that information which is common to all of the sources. This so called shared-GPLVM was further extended by Damianou et al. (2012). Instead of a 'hard' division of the latent space where a latent variable is either shared or private the authors implemented a 'soft' continuous division. This was done by expressing all datasets in terms of all the latent variables and then introducing for each dataset an additional parameter for each latent variable that controls how 'relevant' that latent variable is for that particular data source. A variational approach was used in that model.

Another research direction expanded the GPLVM to include 'output' data (where the outputs could be some continuous random variable or a class label for example). The GPLVM in its original form is an unsupervised method. That is, any additional output information is ignored. The discriminative-GPLVM (Urtasun and Darrell, 2007) incorporates class labels by assuming a prior over the latent variables that favours maximising the between-class variance and minimising the within-class variance. The authors subsequently used Gaussian process classification (GPC) to predict class labels for new individuals. In addition, the kernel matrix from the discriminative-GPLVM could simply be inserted into the GPC prior since the kernel matrix contains all of the information learned about the covariates and the class labels. The discriminative-GPLVM was recently extended by Eleftheriadis et al. (2013) to include multiple datasets. The authors applied their model to facial expression recognition tasks where multiple images of a subjects face are taken from different angles that provide complementary views of the same facial expression. The supervised-GPLVM developed by Gao et al. (2011) allows for continuous output variables to be included. This was done by assuming the output variables were related to the latent variables by some function and then assuming a GP prior over that function. In practise this is not dissimilar to the shared-GPLVM approach since in this case the outputs are essentially treated as another dataset.

Our main advance in this chapter is to develop the Laplace approximation for the marginal likelihood of the GPLVM. The marginal likelihood involves integrating over the latent variables

which is, in general, analytically and numerically intractable. The variational approaches discussed above solve this problem but it is not straightforward to extend such methods to more complicated likelihood functions with non-Gaussian priors or terms that include survival data. The Laplace approximation on the other hand can be easily extended to handle these cases and we will take advantage of this in the following chapter when we combine the GPLVM with a Weibull proportional hazards model. The approximated marginal likelihood is used for the purposes of hyperparameter optimisation and model comparison. We allow for the possibility of multiple datasets by expressing them in terms of a shared set of latent variables as in Shon et al. (2006) and Ek et al. (2008).

This chapter is laid out as follows. In the next section we define the model, describe how to infer the latent variables and hyperparameter, how to make predictions for new individuals, and discuss some of the issues that arise in constructing the Laplace approximation of the marginal likelihood. We then test the model by generating simulated datasets of a particular form that allows us to quantify the accuracy of the model. We explore some of the difficulties in optimising non-linear likelihood functions. Finally, we conduct a binary classification experiment in both high and low dimensional spaces in order to illustrate how reducing the dimension diminishes the symptoms of overfitting (and increases the predictive accuracy).

5.2 The GPLVM

In this section we define the GPLVM and construct the Laplace approximation of the posterior marginal likelihood. Some problems arise with this approximation due to symmetries in the latent variable space which we discus. We then explain how to determine the hyperparameters, and find the optimal embedding in the latent variable space for new individuals.

5.2.1 Model definition

In the most general case we consider S observed datasets $\mathbf{Y}_1, \ldots, \mathbf{Y}_S$ each with N rows and $d_1, \ldots, d_S$ columns respectively. Each row corresponds to one individual and it is assumed that each dataset contains different observations corresponding to the same individuals. For example, we may have gene expression measurements in one dataset and imaging parameters in a second. We assume that each individual i can be represented by a low dimensional vector of latent variables $\mathbf{x}_i \in \mathbb{R}^q$ where $q < \min_s(d_s)$. We assume that covariate μ from source s is

related to the latent variables via

$$y^s_{i\mu} = f^s(\mathbf{x}_i) + \xi^s_{i\mu} \quad \text{for } i = 1, \ldots, N \tag{5.1}$$

where $f^s(\mathbf{x}_i)$ is an unspecified function of the latent variables and $\xi^s_{i\mu} \sim \mathcal{N}(0, \beta^2_s)$ is a noise random variable. Zero mean GP priors with kernel matrices $\mathbf{K}_s$ are then assumed for the functions f^s and it follows that the data likelihood is

$$p(\{\mathbf{Y}_s\}|\mathbf{X}, \{\beta^2_s\}, \boldsymbol{\theta}) = \prod_{s=1}^{S} \prod_{\mu=1}^{d_s} \frac{e^{-\frac{1}{2}\mathbf{y}^s_{:,\mu} \cdot \mathbf{K}_s^{-1} \mathbf{y}^s_{:,\mu}}}{(2\pi)^{N/2}|\mathbf{K}_s|^{1/2}}, \tag{5.2}$$

where $\mathbf{y}^s_{:,\mu}$ denotes the μth column of the matrix $\mathbf{Y}_s$ and the kernel matrix $\mathbf{K}^s_{ij} = k^s(\mathbf{x}_i, \mathbf{x}_j) + \beta^2_s \delta_{ij}$. We have denoted the set of all observed datasets as $\{\mathbf{Y}_s\} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_S\}$ and similarly $\{\beta^2_s\}$ contains all of the noise variance hyperparameters. The vector $\boldsymbol{\theta}$ contains any hyperparameters associated with the kernel function k^s. The data likelihood (5.2) is a product of d_s Gaussian processes for each data source that map the latent variables to the observed data space. Each source is allowed to have different kernel functions that characterise the (possibly non-linear) nature of the mapping. This allows us to simultaneously embed multiple datasets in the same latent variable space where each data source may have a qualitatively different relationship to the latent variables.

5.2.2 Inference of latent variables and hyperparmeters

Within the Bayesian formalism we specify a hierarchy of quantities which we wish to infer from the observed data:

1. The data generating parameters which in our case are the latent variables $\mathbf{X}$. These parameters scale with the number of individuals N.
2. Hyperparameters controlling qualitative features of the model such as the noise levels $\{\beta^2_s\}$ and kernel parameters $\boldsymbol{\theta}$.
3. Models which in our case consist of a choice of kernel function and the dimension of the latent variable space, q.

Using Bayes' theorem we can write the posterior over latent variables as

$$p(\mathbf{X}|\{\mathbf{Y}_s\}, \{\beta^2_s\}, \boldsymbol{\theta}) = \frac{p(\{\mathbf{Y}_s\}|\mathbf{X}, \{\beta^2_s\}, \boldsymbol{\theta})p(\mathbf{X})}{\int \mathrm{d}\mathbf{X}' \, p(\{\mathbf{Y}_s\}|\mathbf{X}', \{\beta^2_s\}, \boldsymbol{\theta})p(\mathbf{X}')} \tag{5.3}$$

with the data likelihood term given by (5.2). We define the negative log likelihood as

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= -\frac{1}{N}\log p(\mathbf{X}|\{\mathbf{Y}_s\},\{\beta_s^2\},\boldsymbol{\theta}) \\ &= \sum_{s=1}^{S}\left(\frac{d_s}{2N}\mathrm{tr}(\mathbf{K}_s^{-1}\mathbf{S}_s)+\frac{d_s}{2N}\log|\mathbf{K}_s|+\frac{d_s}{2}\log 2\pi\right)-\frac{1}{N}\log p(\mathbf{X}) \qquad (5.4)\end{aligned}$$

where $\mathbf{S}_s = d_s^{-1}\mathbf{Y}_s\mathbf{Y}_s^T$. The choice of prior is discussed in Section 5.2.5. This function is numerically minimised with respect to $\mathbf{X}$ using a gradient based optimisation algorithm. Partial derivatives are given in Appendix B.1.

Following (2.35) the posterior over hyperparameters is

$$p(\{\beta_s^2\},\boldsymbol{\theta}|\{\mathbf{Y}_s\}) = \frac{p(\{\mathbf{Y}_s\}|\{\beta_s^2\},\boldsymbol{\theta})p(\{\beta_s^2\})p(\boldsymbol{\theta})}{\int \mathrm{d}\{\beta_s'^2\}\mathrm{d}\boldsymbol{\theta}'\, p(\{\mathbf{Y}_s\}|\{\beta_s'^2\},\boldsymbol{\theta}')p(\{\beta_s'^2\})p(\boldsymbol{\theta}')} \qquad (5.5)$$

where

$$p(\{\mathbf{Y}_s\}|\{\beta_s^2\},\boldsymbol{\theta}) = \int \mathrm{d}\mathbf{X}\, p(\{\mathbf{Y}_s\}|\mathbf{X},\{\beta_s^2\},\boldsymbol{\theta})p(\mathbf{X}). \qquad (5.6)$$

This integral is in general analytically and numerically intractable and so we construct the Laplace approximation as described in Section 2.2.1. From (2.43) we can write

$$p(\{\mathbf{Y}_s\}|\{\beta_s^2\},\boldsymbol{\theta}) \approx p(\{\mathbf{Y}_s\}|\hat{\mathbf{X}},\{\beta_s^2\},\boldsymbol{\theta})(2\pi)^{Nq/2}|(N\mathbf{H})^{-1}(\{\beta_s^2\},\boldsymbol{\theta})|^{1/2} \qquad (5.7)$$

where $\hat{\mathbf{X}} = \min_{\mathbf{X}}\mathcal{L}(\mathbf{X})$. The matrix $\mathbf{H}$ is defined by

$$\mathbf{H}_{p\nu,r\mu} = \frac{\partial}{\partial x_{p\nu}\partial x_{r\mu}}\mathcal{L}(\mathbf{X}), \qquad (5.8)$$

and is calculated explicitly for the three different types of kernels (2.49) in Appendix B.1. The negative log hyperparameter likelihood is defined by

$$\begin{aligned}\mathcal{L}_{hyp}(\{\beta_s^2\},\boldsymbol{\theta}) &= -\frac{1}{N}\log p(\{\beta_s^2\},\boldsymbol{\theta}|\{\mathbf{Y}_s\}) \\ &= \mathcal{L}(\hat{\mathbf{X}}) - \frac{q}{2}\log 2\pi + \frac{1}{2N}\log|N\mathbf{H}(\{\beta_s^2\},\boldsymbol{\theta})|. \qquad (5.9)\end{aligned}$$

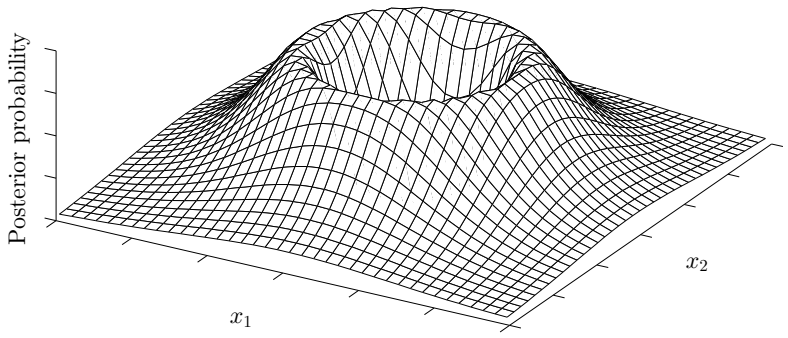


Figure 5.1: Illustration of invariance under unitary transformations. The posterior probability density (5.3) from a toy dataset with $N = 1$ and $\mathbf{x}_1 = (x_1, x_2)$ is plotted (on an arbitrary scale). A vector of observed data $\mathbf{y}_1 \in \mathbb{R}^5$ was randomly generated and the noise level was set to $\beta = 0.01$. The fact that the data likelihood is invariant under rotations and reflections of $\mathbf{x}_1$ through the origin is readily apparent in this case. The points of maximum probability form a circle about the origin. The optimal value of $\mathbf{x}_1$ that is reported by the optimisation algorithm will depend on the initial conditions (which are generated randomly). If either x_1 or x_2 are close to zero then one of the second order derivatives will be close to zero. This renders a Gaussian approximation of the posterior invalid since the covariance matrix is no longer positive definite. Once the rotation and reflection symmetries have been eliminated the posterior reduces to a one dimensional unimodal distribution.

5.2.3 Invariance under unitary transformations

A problem arises during the Laplace approximation due to fact that in the Nq-dimensional posterior search space of latent variables there exist directions in which the second order partial derivatives are zero. These directions point along lines where the log likelihood is constant. This occurs due to the invariance of the log likelihood under rotation or reflection of the latent variables.

To see this let $\mathbf{U}$ be a unitary matrix (corresponding to a rotation or reflection), such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$, and let $\tilde{\mathbf{x}} = \mathbf{U}\mathbf{x}$. Then

$$\tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_j = \mathbf{x}_i \mathbf{U}^{\mathrm{T}} \mathbf{U} \mathbf{x}_j = \mathbf{x}_i \cdot \mathbf{x}_j \text{ and} \tag{5.10}$$

$$(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^2 = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{U}^{\mathrm{T}}\mathbf{U}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^2. \tag{5.11}$$

All of the kernels considered in this paper depend on the covariates solely through expressions of the form (5.10, 5.11) and consequently are invariant under unitary transformations. Since the log likelihood depends on the latent variables via the kernel function it too is invariant under unitary transformations.

There are two undesirable consequences of this property. Firstly, the second order partial derivatives of (5.4) may evaluate to zero and hence $\mathbf{H}$ will not be positive definite and the Gaussian approximation of the marginal likelihood will no longer be well-defined. Secondly, it means that there is not a unique latent variable representation of a dataset but rather an infinite number of mutually equivalent solutions corresponding to different rotations and reflections.

A computationally straightforward solution to this problem is to 'pin down' the latent variable representation such that the symmetries are eliminated. Assuming that we are working in the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$ we demand that $\mathbf{x}_1$ is 'pinned' to the $\mathbf{e}_1$-axis which can always be achieved through an appropriate unitary transformation. We then require the second individual to be confined to the $\mathbf{e}_1$–$\mathbf{e}_2$ plane. This continues for the first $q-1$ individuals. In practice this implies that we simply populate the $(q-1)(q-2)/2$ elements in the upper right hand triangle of $\mathbf{X}$ with zeros and optimise with respect to the remaining latent

variables:

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{x}_{11} & 0 & 0 & 0 \\ \tilde{x}_{21} & \tilde{x}_{22} & 0 & 0 \\ \tilde{x}_{31} & \tilde{x}_{32} & \tilde{x}_{33} & 0 \\ \tilde{x}_{41} & \tilde{x}_{42} & \tilde{x}_{43} & \tilde{x}_{44} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{5.12}$$

To eliminate reflection symmetries we require $\tilde{x}_{11} \geq 0, \tilde{x}_{22} \geq 0, \ldots, \tilde{x}_{qq} \geq 0$. Reflection symmetries do not lead to a problem with zero second order partial derivatives, but to obtain a unique solution it may be desirable to impose the above non-negativity conditions.

Note that the above solution may fail to guarantee a unique solution if $|\mathbf{x}_1| \approx 0$ since 'pinning' a zero vector to the $\mathbf{e}_1$-axis will not constrain the remaining latent variables. Furthermore, if $x_{22} \approx 0$ then reflection symmetry may not be broken. Although the solution may no longer be unique in these cases, the problem of zero partial derivatives will still be avoided.

5.2.4 Making predictions

When we observe a new sample, $\mathbf{y}^*$, we wish to firstly infer its optimal embedding, $\mathbf{x}^*$, in the latent variable space. We use the GP predictive distribution $p(\mathbf{y}^*|\mathbf{x}^*) \sim \mathcal{N}(\mathbf{m}, \kappa^2)$ from (2.51, 2.52) with

$$m_\mu = \mathbf{k} \cdot \mathbf{K}^{-1} \mathbf{y}_\mu \tag{5.13}$$

$$\kappa^2 = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k} \cdot \mathbf{K}^{-1} \mathbf{k} + \beta^2. \tag{5.14}$$

In general we will have several data sources and the posterior is given by

$$p(\mathbf{x}^*|\mathbf{y}_1^*, \ldots \mathbf{y}_S^*) \propto p(\mathbf{y}_1^*|\mathbf{x}^*) \cdots p(\mathbf{y}_S^*|\mathbf{x}^*) p(\mathbf{x}^*). \tag{5.15}$$

Observations are not necessarily required from all of the data sources since we can simply omit terms in (5.15) corresponding to unobserved data. The negative log likelihood of the predictive distribution is minimised numerically with respect to $\mathbf{x}^*$ using gradient based methods (partial

derivatives are given in Appendix B.2):

$$\begin{aligned}\mathcal{L}(\mathbf{x}^*) &= -\frac{1}{N}\log p(\mathbf{x}^*|\mathbf{y}_1^*,\ldots,\mathbf{y}_S^*) \\ &= \frac{1}{N}\sum_{s=1}^{S}\left(\frac{1}{2\kappa_s^2}(\mathbf{y}_s^*-\mathbf{m}_s)^2+\frac{d_s}{2}\log 2\pi+d_s\log\kappa_s\right). \end{aligned} \tag{5.16}$$

Examples of the negative log predictive likelihood are plotted in Figure 5.3.

5.2.5 Implementation

In this section we give a detailed account of how the model is used in practice. Once presented with a dataset $\mathbf{Y}$ there are a number of goals we wish to achieve with the model:

1. Extraction of latent variables $\mathbf{X}$.
2. Inference of the hyperparameters.
3. Both of the above in the case of multiple datasets $\{\mathbf{Y}_s\}$.
4. Determining the most appropriate value of q and kernel function.

We will now explain how to perform each one of these tasks separately.

1. Extraction of latent variables X

We assume for the purposes of this section that the choice of q and kernel type have been made and that the values of the hyperparameters are fixed.

1. Pick random initial values for the latent variables[1] $\mathbf{X}$.
2. Numerically solve $\hat{\mathbf{X}} = \min_{\mathbf{X}} \mathcal{L}(\mathbf{X})$ using gradient based optimisation. In the case of non-linear kernels $\mathcal{L}(\mathbf{X})$ will possess multiple local minima in general. Consequently, multiple attempts are made to locate the global minimum. Each attempt starts from a different random starting point. Computationally, these attempts can be made in parallel.

[1] In the original implementation of the GPLVM the matrix $\mathbf{X}$ was initialised to the first q principal components which is computationally faster than conducting multiple searches to find the global minimum, but may lead to a suboptimal solution if the initial search point leads to a local minimum.

2. Inference of the hyperparameters.

To determine the optimal values for hyperparameters we numerically solve min $\mathcal{L}_{hyp}(\{\beta_s^2\}, \boldsymbol{\theta})$ (from (5.9)) using either the Nelder-Mead search algorithm, or a line search algorithm if there is only one hyperparameter. Each evaluation of $\mathcal{L}_{hyp}(\{\beta_s^2\}, \boldsymbol{\theta})$ requires $\hat{\mathbf{X}}$ to be recalculated since the location of the minimum may change for different values of hyperparameters. However, the value of $\mathbf{X}$ can be initialised to its previously optimal value since a small change in hyperparameters will not induce a large change in $\mathbf{X}$ in general.

3. Dealing with multiple datasets

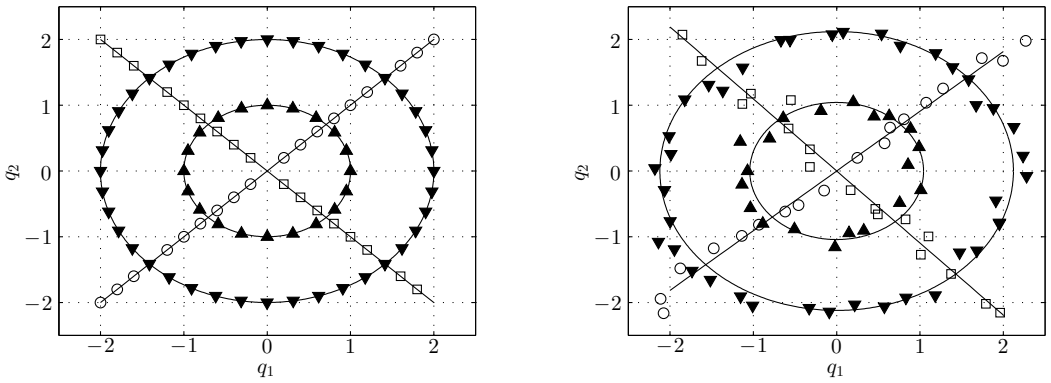
When we observe S datasets $\mathbf{Y}_1, \ldots, \mathbf{Y}_S$ we begin by finding the optimal hyperparameters for each source separately. We then minimise (5.4) with respect to $\mathbf{X}$, while holding the hyperparameters for each source fixed.

4. Determining the most appropriate value of q and kernel function.

In order to do this we simply chose a value of q and locate the minimum of the negative log hyperparameter likelihood (5.9). We then compare the minimum value of (5.9) for different values of q to see which is most probable. Within the Bayesian formalism we should determine q after integrating over the hyperparameters (as in (2.36)) but this is not feasible here. Using the value of the hyperparameter posterior is an approximation, but one that appears to be satisfactory in this case.

Choice of priors

The posterior density for the latent variables (5.3) requires a choice of prior for $\mathbf{X}$. One possibility is $p(\mathbf{X}|\sigma_1^2) = \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ with the hyperparameter σ_1^2 controlling the prior variance. Now the kernel parameter σ from (2.49), the hyperparameter σ_1 and the overall length scale of $\mathbf{X}$ influence the variance of the GP prior however. In other words, when we attempt to find the optimal solution there are several routes that the model can take to specify the overall variance of the high dimensional covariates. This redundancy is undesirable since the parameters and hyperparameters may not be uniquely determined by the observed data. For the linear and polynomial kernel we therefore set $\sigma = 1$ with a flat prior over $\mathbf{X}$. Observed data are normalised to zero mean and unit variance. The overall length scale of $\mathbf{X}$ is thus naturally determined by $\mathbf{Y}$ without requiring additional hyperparameters. For the squared



(a) 'True' latent variables

(b) Retrieved latent variables

Figure 5.2: On the left are the 'true' latent variables that are projected into a high dimensional space to produce simulated high dimensional data with $d = 10$ and a linear kernel. Gaussian noise with variance $\beta^2 = 0.1$ is added. The data are arranged in a particular geometric pattern that allows us to measure how well the model retrieves the original latent variables. On the right are the latent variables retrieved by the GPLVM from the high dimensional data. Misallignment errors, defined in the text, are $\mathcal{E}_{radial} = 0.0051$, $\mathcal{E}_{angular} = 0.0086$ and $\mathcal{E}_{linear} = 0.0288$.

exponential kernel this is not the case because the length scale of $\mathbf{X}$ cannot be used to change the magnitude of terms in the kernel matrix. Hence we optimise over σ and use $p(\mathbf{X}|\sigma_1^2) = \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$.

5.3 Results

In this section we ran a number of simulation studies and examined the performance of the model under various conditions. We propose a somewhat ad hoc method of quantitatively assessing the performance of the model. We ran simulation studies to see if the model is capable of retrieving the correct low dimensional structure. Finally, we explore the effects of overfitting by using a Support Vector Machine (SVM) (Vapnik, 1995) to perform a binary classification task in both high and low dimensional spaces. Before we discus the results we first describe how the simulated data are generated. More thorough investigations will be postponed until the next chapter.

5.3.1 Generation of simulated data

To generate simulated data we begin by picking latent variables either at random or manually. These latent variables are then projected into a high dimensional space. The high dimensional covariates are the outputs of a Gaussian process on $\mathbf{X}$. This means that in practice we first compute the GP kernel matrix $\mathbf{K}$, and secondly draw N-dimensional vectors from a multivariate Gaussian with covariance matrix $\mathbf{K}$. Each vector drawn corresponds to one covariate so this can simply be repeated to build up a high dimensional dataset of arbitrary dimension.

It will be helpful to compare the retrieved $\hat{\mathbf{X}}$ with the 'true' values $\mathbf{X}$. For this purpose we choose the specific latent variables plotted in Figure 5.2 (a) which are arranged in a specific geometrical pattern that allow quantitative measures of similarity to be defined. The samples that belong to either of the two circles, for instance, should be equidistant from the origin. If $\tilde{r}$ is the mean distance from the origin then we can define the radial error as

$$\mathcal{E}_{radial} = \frac{1}{|C|} \sum_{i \in C} \frac{|\mathbf{x}_i| - \tilde{r}}{\tilde{r}} \tag{5.17}$$

where C is the set of points belonging to the circle and $|C|$ is the number of samples belonging to that set. The error for both circles are averaged.

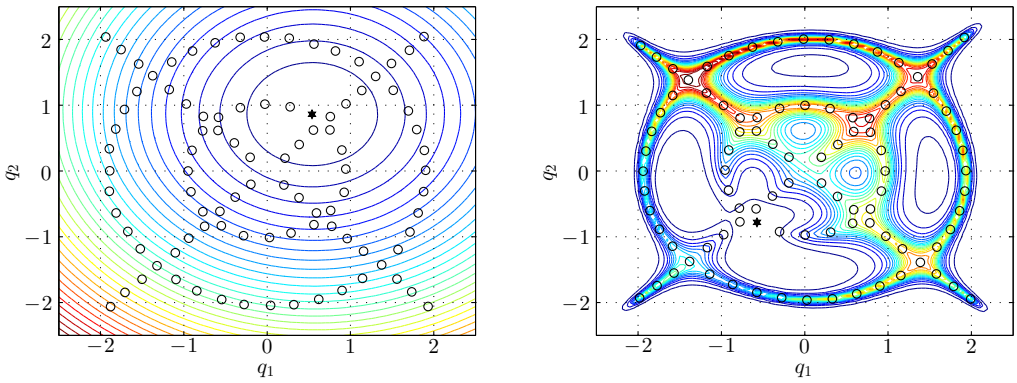
Similarly, the angles between each pair of samples belonging to each circle should be equal. In the case of the larger circle the angular separation should be $\tilde{\theta} = 2\pi/20$. If we let $\Delta\theta_i$ denote the angle between $\mathbf{x}_i$ and the neighbouring point then we can define the mean angular error as

$$\mathcal{E}_{angular} = \frac{1}{|C|} \sum_{i \in C} \frac{\Delta\theta_i - \tilde{\theta}}{\tilde{\theta}}. \tag{5.18}$$

For both of the lines we can attempt a linear fit by writing $q_2 = \alpha q_1$. The value of α which minimises the sum of squared errors $\sum_i (x_{i2} - \alpha x_{i1})^2$ is given by $\hat{\alpha} = \sum x_{i1} x_{i2} / \sum x_{i1}^2$. We can then define the total sum of squares $SS_{tot} = \sum (x_{i2} - \langle x_{i2} \rangle)^2$ and the sum of squared residuals $SS_{err} = \sum (x_{i2} - \hat{\alpha} x_{i1})^2$ and finally define

$$\mathcal{E}_{linear} = \frac{SS_{err}}{SS_{tot}} \tag{5.19}$$

Note that $1 - \mathcal{E}_{linear}$ is called the coefficient of determination and is typically denoted by R^2 and takes a value between 0 and 1 where 1 corresponds to a perfect linear fit. These



(a) Linear kernel

(b) Squared exponential kernel

Figure 5.3: Contour plots of the negative log predictive likelihood for both a linear ($d = 100$, $\beta^2 = 0.01$) and squared exponential ($d = 100$, $\beta^2 = 0.01$, $\sigma = 1$, $l = 1$) kernel. The dataset from Figure 5.2 (a) was used with one individual omitted while the model was trained. The inferred positions of the training individuals are denoted with white circles. The global minimum of the predictive distribution for the omitted individual is denoted with the black star. In both cases the model is able to correctly predict the location of the omitted individual in the latent space. Note the in the linear case there is only one minimum whereas in the non-linear case there are multiple local minima. As a result multiple attempts must be made to locate the global minimum. Finally, note that despite the fact that the same individual was removed in both cases the solution on the left is a reflection (through both the q_1 and q_2 axes) of the solution on the right. This is due to the symmetries discussed in Section 5.2.3.

misalignment error measures have two desirable properties. Firstly, all three errors are zero for the 'true' latent variables. Secondly, the error measures are invariant under rescaling of $\mathbf{X}$.

5.3.2 Example of local minima in the negative log predictive likelihood

In Figure 5.3 we generated simulated data using the latent variables from Figure 5.2 (a). This time we removed one of the individuals however, and once we had trained the GPLVM we used it to predict the location of the removed individual. We did this for both a linear and a squared exponential kernel. In both cases the model was able to accurately predict the location of that individual in the latent variable space. In the case of a linear kernel the negative log predictive likelihood function appears to be unimodal and therefore straightforward to optimise. In the case of the squared exponential kernel there are multiple local minima and consequently multiple attempts are required to determine the global minimum.

5.3.3 Reducing the effects of overfitting in a binary classification task

To illustrate the practical benefit of a latent variable representation in reducing overfitting we now generate low dimensional data $\mathbf{X}$ with $q = 2$ where each sample belongs to a binary class $\{-1, +1\}$. We choose $N = 200$, $d = 100$ and a comparatively high noise level of $\beta^2 = 8.0$. We generate 100 samples with a class label of $+1$ from a Gaussian distribution with unit variance and mean $(1, 1)$. We then generate 100 samples from the -1 class from two unit variance Gaussians centred on $(-\frac{1}{2}, -\frac{1}{2})$ and $(\frac{1}{2}, -\frac{1}{2})$. Once we have generated $\mathbf{X}$ we project these data into a high dimensional space using a linear kernel as described above.

We used a Support Vector Machine (SVM) (Vapnik, 1995) to perform this binary classification task. An SVM relates a vector of input variables to a class label $\{+1, -1\}$. The SVM requires a kernel function to be specified, in a similar manner to GP regression. The kernel function roughly tells us how 'similar' any two inputs are. If we use a linear kernel then the SVM will find the optimal linear separating hyperplane that separates both classes. In our simulated data here we will use a squared exponential kernel since we do not expect these data to be linearly separable.[2]

The SVM itself depends on some parameters which must be tuned. In our case these are the so-called 'box constraint' parameter (which controls the degree to which the SVM will try and avoid misclassifications during training) and the length scale in the squared exponential kernel. We use leave-one-out cross validation (LOOCV) to determine the optimal parameter values. This proceeds by removing one of the training samples and training the SVM on the remaining samples. The trained SVM is then used to predict the class label of the removed sample. This is repeated until each sample in the training set has been removed once. The total proportion of correct predictions is then computed. The SVM parameters are adjusted so as to maximise this number. In other words, the SVM is made to fit the training data as well as possible.

In our application the simulated high dimensional data are split into a training set, $\mathbf{Y}_{tr} \in \mathbb{R}^{100\times100}$ and a validation set, $\mathbf{Y}_{val} \in \mathbb{R}^{100\times100}$, of equal size. The GPLVM is used to generate a two dimensional representation, $\mathbf{X}^*_{tr} \in \mathbb{R}^{100\times2}$, from the training set. We then predict the optimal low dimensional embedding for each individual in the high dimensional validation set to construct $\mathbf{X}^*_{val} \in \mathbb{R}^{100\times2}$, as described in Section 5.2.4.

[2]The kernel function of the SVM should not be confused with the kernel function used in the GPLVM. In this case the SVM uses a squared exponential kernel (for non-linear binary predictions) and the GPLVM uses a linear kernel (for mapping the latent variables to the high dimensional covariates) and these are chosen independently.

Once this has been done SVMs are then trained on both the high dimensional data $\mathbf{Y}_{tr}$ and the latent variable representation $\mathbf{X}^*_{tr}$ and the training success rate is computed. This is the percentage of individuals in the training set that are correctly classified by the trained SVM. We then use these trained SVMs to make predictions for each individual in both the high and low dimensional validation sets and compute the validation success rates. The entire experiment is then repeated fifty times and the results averaged.

The results are shown in Table 5.1. The most important figures are the validation success rates since these quantify the extent to which the SVM can generalise to unseen individuals. When we use the latent variable representation we achieve a success rate of 87.9% compared to 82.3% when we use the original high dimensional data. This increase in predictive accuracy is because the effect of overfitting is diminished by reducing the dimension. In the 100 dimensional space the SVM is unable to extract genuine patterns as easily it can in a 2 dimensional space.

	High dimensional data ($d = 100$)	Latent variables ($q = 2$)
Mean training success	$92.0 \pm 7.3\%$	$91.9 \pm 3.3\%$
Mean validation success	$82.3 \pm 4.3\%$	$87.9 \pm 3.3\%$

Table 5.1: Results from running an SVM binary classifier with a squared exponential kernel on the original 100-dimensional dataset and a 2-dimensional latent variable representation. The percentages are the mean percentage of correct binary classifications using the SVM. The entire experiment was repeated fifty times and we have included plus and minus one standard deviation calculated over these repetitions. The data have been split into equal sized training and validation sets. We can see that the success rate (which is the average number of correctly predicted class labels) is higher in the latent variable space. The degradation of predictive performance in the high dimensional space is a hallmark of overfitting; the SVM struggles to detect meaningful patterns in an overwhelming volume of data.

5.4 Discussion

We have explored the use of the GPLVM as a tool to tackle high dimensional data with an emphasis on increasing predictive accuracy by reducing the detrimental effects of overfitting. By decreasing the ratio of covariates to patients we aimed to extract patterns from the data more robustly. The main technical advance in this chapter was to construct the Laplace approximation of the marginal likelihood. This will be useful in the following chapter when the GPLVM is combined with the Weibull proportional hazards model since we only require the

second order partial derivatives of the log likelihood (which can easily be obtained). Although variational approaches to approximating the log likelihood have been developed elsewhere they cannot readily be applied to a model with terms from a WPHM without considerable effort. The Laplace approximation therefore gives us the freedom to easily experiment with alternative likelihood functions provided we can obtain the second order derivatives.

We also eliminated various symmetries than exist in the latent variable search space. This was achieved in a computationally elegant manner and has two attractive consequences. The first is that the Laplace approximation is guaranteed to be well-defined, and the second is that the latent variable solution is unique.

We began to explore the performance of the model by generating a specific type of simulated high dimensional data that allowed for quantitative assessment of how accurate the recovered latent variables are. Simulation studies of a binary classification task performed in both high and low dimensional spaces illustrate the practical benefit the GPLVM has in terms of combating the effects of overfitting. We found that the predictive accuracy increases after we reduced the dimension of the high dimensional data. Further investigation of the model will be undertaken in the next chapter.

Chapter 6

Simultaneous dimensionality reduction with survival analysis

6.1 Introduction

We now build on the work of the previous chapter and combine the Gaussian process latent variable model (GPLVM) with a Weibull proportional hazards model (WPHM) to create a method for dealing with high dimensional survival data. As discussed in the previous chapter high dimensional data consist of a large number of covariates d and comparatively few individuals N. This leads to the problem of overfitting where it becomes difficult to extract meaningful relationships between such large quantities of data. The phenomenon of overfitting is also problematic for survival data. For example, fitting a Cox model when $d > N$ is difficult as the regression coefficients will not be unique (Witten and Tibshirani, 2010), a situation that is similar to simple linear regression.

There are a variety of strategies for tackling high dimensional data. They can broadly be divided into supervised methods (which use outcome information) and unsupervised methods (which search for structure solely within the covariates and ignore survival outcomes). A popular supervised method is univariate feature selection. For example, a Cox proportional hazards model can be fitted for each covariate separately and then the covariates can be ranked according to some measure of statistical significance. This still suffers from the risk of overfitting but it at least allows researchers to extract a more manageable subset of covariates.

An alternative strategy is to impose some form of regularisation on a regression model. In ridge regression an L_2 penalty $\chi \sum b_i^2$ is added to the negative log likelihood, where $\mathbf{b}$ is

the vector of regression coefficients. This encourages most of the regression coefficients to be small and helps alleviate overfitting since more complicated regression models will be avoided. In the Bayesian formalism this is equivalent to assuming a Gaussian prior over $\mathbf{b}$ and can be interpreted as prior knowledge that we expect only a subset of the covariates to be relevant. Alternatively an L_1 penalty $\chi \sum |b_i|$ can be used which results in a sparse solution as most of the regression coefficients shrink to zero. This is also known as lasso regression and both types of regularisation schemes have been applied to survival data (Verweij and Van Houwelingen, 1994; Tibshirani, 1997). See also Goeman (2010), Park and Hastie (2007) and Sohn et al. (2009) for examples of L_1 penalised Cox regression. In addition, Engler and Li (2009) used elastic net regression for variable selection in a Cox model and Ishwaran et al. (2011) adapted random forest variable selection to survival data.

Unsupervised methods offer yet another route to analysing high dimensional data that do not take into account outcome information. One popular approach is hierarchical clustering which groups individuals into different clusters. This can be used to identify different subtypes of a disease for example, or to look for subgroups with different survival characteristics. Another approach is dimensionality reduction techniques which try to find a low dimensional representation of high dimensional data. The most popular method is arguably Principal Component Analysis (PCA) which finds the directions of maximum variance (the principal components) in the data space. One can then select the top q principal components and thereby achieve a dimensionality reduction from d covariates to q principal components. A method that combines PCA with survival outcomes was developed Bair and Tibshirani (2004) and Bair et al. (2006), which they call supervised principal components. The GPLVM discussed in the last chapter can be considered a non-linear generalisation of PCA. It possesses a number of additional attractive features such as the ability to combine multiple data sources and to probabilistically determine the optimal value of q, the number of latent variables. In this chapter we will extend the GPLVM to incorporate survival data. An excellent review of survival analysis methods for high dimensional data can be found in Witten and Tibshirani (2010).

It is worth noting that the strategy pursued to analyse survival data will depend on the type of question that is being asked. If the goal is to establish associations between covariates and survival outcomes then feature selection methods or regularised regression techniques would be appropriate since they will attempt to pick out the relevant covariates. If however the goal is to make accurate predictions then dimensionality reduction methods are desirable since they utilise all information while trying to diminish the risk of overfitting. However,

dimensionality reduction methods generally make it difficult to interpret what impact a certain covariate has on the outcome because the low dimensional variables are a (potentially non-linear) combination of the high dimensional covariates. In this work we place more emphasis on making accurate predictions. By reducing the dimension we hope to reduce the effects of overfitting and consequently increase our predictive accuracy.

The combined GPLVM-WPHM that we will develop in this chapter is a supervised dimensionality reduction method since it now includes the survival outcomes. A supervised method is desirable since there is no guarantee that an unsupervised low dimensional representation (where outcome information was ignored) will contain information that is relevant to survival outcomes. By combining both the covariates and survival outcomes we hope to infer an embedding in a low dimensional manifold that captures the relationship between covariates and survival outcomes. The combined model will be forced to compromise between generating latent variables that explain the high dimensional covariates well but that are also consistent with the survival outcomes. By connecting the covariates to event times in the latent variable space we are constraining the degrees of freedom the model has and consequently the risk of overfitting is diminished.

A helpful analogy to our combined model can be found in the recent work of Eleftheriadis et al. (2013). The authors developed a discriminative-GPLVM for the purposes of facial expression recognition. Photographic images of a subject's face are acquired from two different angles and these two images are regarded as two separate data sources (that are expressed in terms of the same latent variables). This is appropriate since both data sources correspond to the same facial expression, just from different perspectives. The authors also include class label information while training the model. We can interpret our combined GPLVM-WPHM in a similar manner. Our application is to multiple sources of data acquired from patients. The latent variables represent the underlying biological processes we are interested in. Different datasets provide different 'perspectives' or 'views' of these underlying processes. For example, gene expression data and biomedical imaging data may provide complementary information on what is happening inside a cancer cell. If the survival outcomes are driven by the same biological processes then these too constitute a third 'perspective' of the latent variable space. By combining all relevant pieces of information we aim to build a clearer characterisation of the underlying cellular processes.

In the following section we define the combined model and explain how to infer the relevant parameters and hyperparameters. We pay particular attention to the choice of Bayesian priors as this is found to impact the model performance significantly. In Section 6.3 we present results

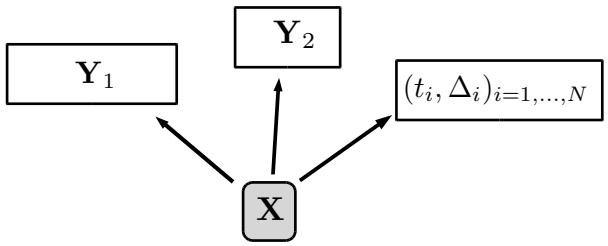


Figure 6.1: Schematic diagram of the combined GPLVM-WPHM. We attempt to express both of the high dimensional covariate datasets $\mathbf{Y}_1 \in \mathbb{R}^{N \times d_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times d_2}$ and the observed survival outcomes (t_i, Δ_i), where $i = 1, \ldots, N$, in terms of the same low dimensional latent variables. We thereby generate a more parsimonious representation of information from all of the available data sources. N is the total number of individuals, and d_1 and d_2 are the number of covariates in the first and second datasets respectively. A key assumption is that each source of information is conditionally independent given the latent variables. By reducing the dimension of the data we aim to alleviate the symptoms of overfitting. Including the survival outcomes during the dimensionality reduction process helps to ensure that the latent variables will capture information that is relevant to survival probabilities.

from simulation studies and some real gene expression data from breast cancer patients.

6.2 Combining the GPLVM and the Weibull proportional hazards model

Most of the groundwork for the combined model has already been covered in Chapter 5 (the GPLVM) and Section 3.4.2 (the WPHM). Our key assumption is that the survival outcomes and high dimensional covariates are conditionally independent given the latent variables. This allows us to write the data likelihood as a product of the GPLVM and WPHM likelihoods. Using Bayes' theorem we optimise the posterior over the latent variables which will now attempt to account for both the observed covariates and the survival outcomes simultaneously. Particular attention is paid to the choice of prior distributions since these were found to be crucial in preventing nonsensical solutions.

6.2.1 Model definition

We assume that the observed data consist of multiple high dimensional covariate measurements $\mathbf{Y}_1, \ldots, \mathbf{Y}_S$ with $\mathbf{Y}_s \in \mathbb{R}^{N \times d_s}$. In addition we also observe survival outcomes, denoted by $D = \{(\tau_1, \Delta_1), \ldots, (\tau_N, \Delta_N)\}$ where $\tau_i > 0$ is the time until the event of interest and

$\Delta_i = 0$ if individual i is right censored and $\Delta_i = 1$ indicates that the event of interest occurred first. We will not consider the case of competing risks in this work and will assume that the time-to-event for right censoring is statistically independent of the time-to-event for the primary risk. To avoid confusion between the noise hyperparameters of the latent variable model we will denote the WPHM regression coefficients as $\mathbf{b}$ (instead of the more common $\boldsymbol{\beta}$ used in earlier chapters). From Bayes' theorem, the posterior density is

$$p(\mathbf{X}, \mathbf{b}, \rho, \nu | \{\mathbf{Y}_s\}, D, \boldsymbol{\theta}, \{\beta_s\}) = \frac{p(\{\mathbf{Y}_s\}, D | \mathbf{X}, \mathbf{b}, \rho, \nu, \boldsymbol{\theta}, \{\beta_s\}) p(\mathbf{X}) p(\mathbf{b}) p(\rho) p(\nu)}{p(\{\mathbf{Y}_s\}, D | \boldsymbol{\theta}, \{\beta_s\})} \tag{6.1}$$

where

$$p(\{\mathbf{Y}_s\}, D | \boldsymbol{\theta}, \{\beta_s\}) = \int \mathrm{d}\mathbf{X}\, \mathrm{d}\mathbf{b}\, \mathrm{d}\rho\, \mathrm{d}\nu\, p(\{\mathbf{Y}_s\}, D | \mathbf{X}, \mathbf{b}, \rho, \nu, \boldsymbol{\theta}, \{\beta_s\}) p(\mathbf{X}) p(\mathbf{b}) p(\rho) p(\nu). \tag{6.2}$$

We now assume conditional independence between the observed covariate data and the survival data given the latent variables:

$$p(\{\mathbf{Y}_s\}, D | \mathbf{X}, \mathbf{b}, \rho, \nu, \boldsymbol{\theta}, \{\beta_s\}) = p(\{\mathbf{Y}_s\} | \mathbf{X}, \boldsymbol{\theta}, \{\beta_s\}) p(D | \mathbf{X}, \mathbf{b}, \rho, \nu, \boldsymbol{\theta}). \tag{6.3}$$

The first term is the data likelihood (5.2) from the GPLVM which is

$$p(\{\mathbf{Y}_s\} | \mathbf{X}, \boldsymbol{\theta}, \{\beta_s^2\}) = \prod_{s=1}^{S} \prod_{\mu=1}^{d_s} \frac{e^{-\frac{1}{2}\mathbf{y}_{:,\mu}^s \cdot \mathbf{K}_s^{-1} \mathbf{y}_{:,\mu}^s}}{(2\pi)^{N/2} |\mathbf{K}_s|^{1/2}}. \tag{6.4}$$

The second term is the data likelihood (3.42) from the WPHM model described in Section 3.4.2. This term is

$$p(D | \mathbf{X}, \mathbf{b}, \rho, \nu, \boldsymbol{\theta}) = \prod_{i=1}^{N} [\lambda_0(\tau_i) e^{\mathbf{b} \cdot \mathbf{x}_i}]^{\Delta_i} \exp(-\Lambda_0(\tau) e^{\mathbf{b} \cdot \mathbf{x}_i}). \tag{6.5}$$

6.2.2 Inference of latent variables, regression parameters and hyperparameters

We then take the negative log of (6.1):

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mathbf{b}, \rho, \nu) &= -\frac{1}{N} \log p(\mathbf{X}, \mathbf{b}, \rho, \nu | \{\mathbf{Y}_s\}, D, \boldsymbol{\theta}, \{\beta_s\}) \\ &= \sum_{s=1}^{S} \left(\frac{d_s}{2N} \mathrm{tr}(\mathbf{K}_s^{-1} \mathbf{S}_s) + \frac{d_s}{2N} \log |\mathbf{K}_s| + \frac{d_s}{2} \log 2\pi \right) \\ &\quad - \frac{1}{N} \sum_{i:\Delta_i=1} (\log \lambda_0(\tau_i) + \mathbf{b} \cdot \mathbf{x}_i) + \frac{1}{N} \sum_{i=1}^{N} \Lambda_0(\tau_i) e^{\mathbf{b} \cdot \mathbf{x}_i} \\ &\quad - \frac{1}{N} \log p(\mathbf{b}) - \frac{1}{N} \log p(\rho) - \frac{1}{N} \log p(\nu). \end{aligned} \tag{6.6}$$

We determine point estimates of the unknown parameters by numerically minimising the function $\mathcal{L}(\mathbf{X}, \mathbf{b}, \rho, \nu)$ with respect to its arguments. This is done by first optimising with respect to $\mathbf{X}$ while $\mathbf{b}$, ρ, and ν are held fixed. We then fix $\mathbf{X}$ to its optimal value and optimise with respect to the WPHM parameters. Optimisation alternates between both sets of parameters until convergence to a stable solution. Further details are given in Section 6.2.4.

The posterior over hyperparameters is

$$p(\{\beta_s^2\}, \boldsymbol{\theta} | \{\mathbf{Y}_s\}, D) = \frac{p(\{\mathbf{Y}_s\}, D | \{\beta_s^2\}, \boldsymbol{\theta}) p(\{\beta_s^2\}) p(\boldsymbol{\theta})}{\int \mathrm{d}\{\beta_s'^2\} \mathrm{d}\boldsymbol{\theta}' \, p(\{\mathbf{Y}_s\}, D | \{\beta_s'^2\}, \boldsymbol{\theta}') p(\{\beta_s'^2\}) p(\boldsymbol{\theta}')}, \tag{6.7}$$

where the marginal density $p(\{\mathbf{Y}_s\}, D | \{\beta_s^2\}, \boldsymbol{\theta})$ is defined by (6.2). Again, this integral is generally intractable both analytically and numerically. We therefore construct the Laplace approximation as described in Section 2.2.1. From (2.43) we can write

$$p(\{\mathbf{Y}_s\}, D | \{\beta_s^2\}, \boldsymbol{\theta}) \approx p(\{\mathbf{Y}_s\}, D | \hat{\mathbf{X}}, \hat{\mathbf{b}}, \hat{\rho}, \hat{\nu}, \boldsymbol{\theta}, \{\beta_s\}) (2\pi)^{Nq/2} |(N\mathbf{H})^{-1}(\{\beta_s^2\}, \boldsymbol{\theta})|^{1/2}. \tag{6.8}$$

The 'hats' denote values that minimise (6.6). The matrix $\mathbf{H}$ is a block matrix of second order partial derivatives:

$$\mathbf{H}(\{\beta_s^2\}, \boldsymbol{\theta}) = \begin{pmatrix} \mathbf{H}_{XX} & \mathbf{H}_{Xb} & \mathbf{h}_{X\rho} & \mathbf{h}_{X\nu} \\ \mathbf{H}_{Xb} & \mathbf{H}_{bb} & \mathbf{h}_{b\rho} & \mathbf{h}_{b\nu} \\ \mathbf{h}_{X\rho} & \mathbf{h}_{b\rho} & h_{\rho\rho} & h_{\rho\nu} \\ \mathbf{h}_{X\nu} & \mathbf{h}_{b\nu} & h_{\rho\nu} & h_{\nu\nu} \end{pmatrix}. \tag{6.9}$$

Partial derivatives involving $\mathbf{X}$ are

$$[\mathbf{H}_{XX}]_{p\eta,r\mu} = \frac{\partial^2 \mathcal{L}}{\partial x_{p\eta} \partial x_{r\mu}}, \qquad [\mathbf{H}_{Xb}]_{p\eta,\mu} = \frac{\partial^2 \mathcal{L}}{\partial x_{p\eta} \partial b_{\mu}}$$
$$[\mathbf{h}_{X\rho}]_{p\eta} = \frac{\partial^2 \mathcal{L}}{\partial x_{p\eta} \partial \rho}, \qquad [\mathbf{h}_{X\nu}]_{p\eta} = \frac{\partial^2 \mathcal{L}}{\partial x_{p\eta} \partial \nu}, \tag{6.10}$$

and are similar to the partial derivatives of the negative log GPLVM likelihood but with additional terms coming from the WPHM component of the combined model. The terms containing the WPHM regression coefficients are

$$[\mathbf{H}_{bb}]_{\eta,\mu} = \frac{\partial^2 \mathcal{L}}{\partial b_{\eta} \partial b_{\mu}}, \qquad [\mathbf{h}_{b\rho}]_{\eta} = \frac{\partial^2 \mathcal{L}}{\partial b_{\eta} \partial \rho}, \qquad [\mathbf{h}_{b\nu}]_{\eta} = \frac{\partial^2 \mathcal{L}}{\partial b_{\eta} \partial \nu}, \tag{6.11}$$

all of which are identical to those obtained from an independent WPHM model. The remaining partial derivatives are also identical to an independent WPHM model:

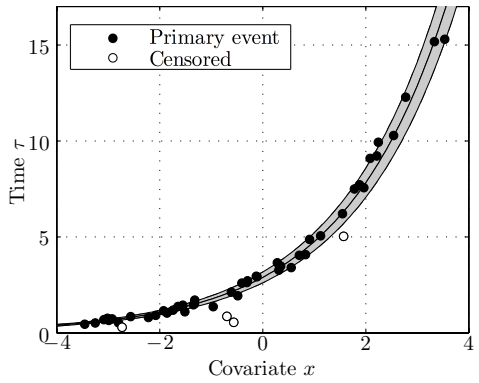
$$h_{\rho\rho} = \frac{\partial^2 \mathcal{L}}{\partial \rho^2}, \qquad h_{\nu\nu} = \frac{\partial^2 \mathcal{L}}{\partial \nu^2}, \qquad h_{\rho\nu} = \frac{\partial^2 \mathcal{L}}{\partial \rho \partial \nu}. \tag{6.12}$$

All of these are calculated explicitly in Appendix B.3. The negative log hyperparameter likelihood is defined by

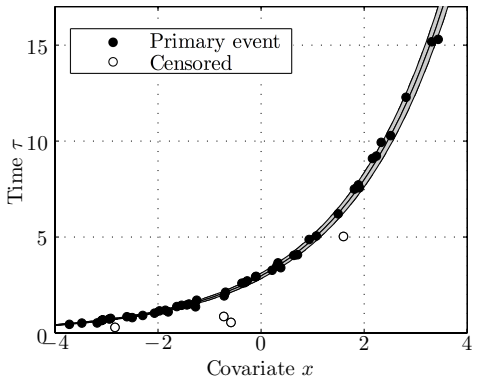
$$\begin{aligned} \mathcal{L}_{hyp}(\{\beta_s^2\}, \boldsymbol{\theta}) &= -\frac{1}{N} \log p(\{\beta_s^2\}, \boldsymbol{\theta} | \{\mathbf{Y}_s\}, D) \\ &= \mathcal{L}(\hat{\mathbf{X}}, \hat{\mathbf{b}}, \hat{\rho}, \hat{\nu}) - \frac{q}{2} \log 2\pi + \frac{1}{2N} \log |N\mathbf{H}(\{\beta_s^2\}, \boldsymbol{\theta})|. \end{aligned} \tag{6.13}$$

6.2.3 The perils of assuming uniform priors

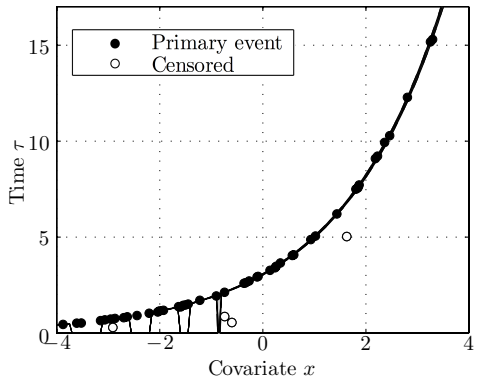
In previous chapters we have assumed flat improper priors for the parameters in the WPHM. This can sometimes lead to spurious results where the negative log likelihood diverges towards minus infinity. These undesirable solutions originate during the optimisation of the latent variables and the parameters of the WPHM. As optimisation alternates between both sets of parameters the latent variables are gradually adjusted such that they begin to 'perfectly fit' the WPHM model. The parameters of the WPHM are subsequently adjusted such that a highly specific model of the survival times is specified — a model which fits the survival data $(\mathbf{x}_i, \tau_i)$ almost perfectly. A highly specific WPHM event time density is achieved with



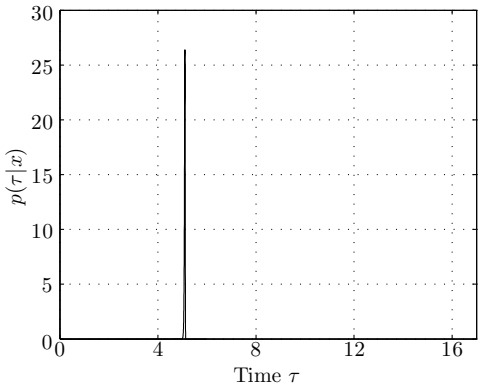
(a) $b = -6.24$, $\rho = 3.01$, $\nu = 12.52$



(b) $b = -13.16$, $\rho = 3.03$, $\nu = 26.63$

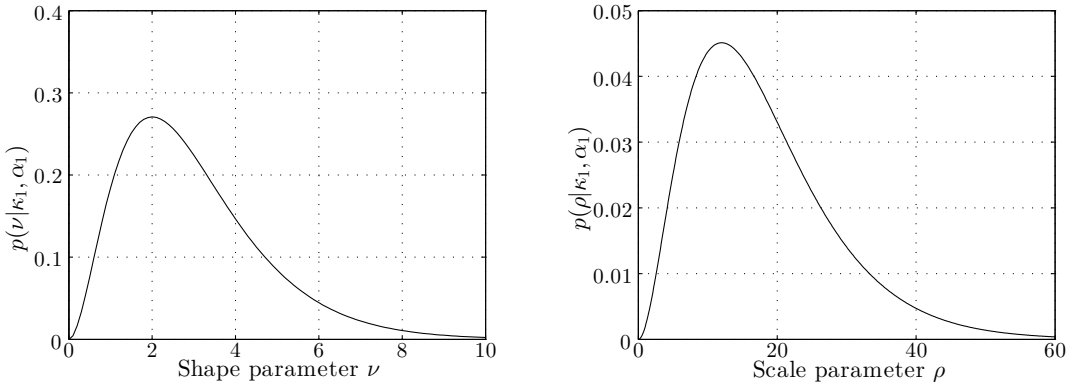


(c) $b = -54.83$, $\rho = 3.07$, $\nu = 111.3$



(d) $b = -180.3$, $\rho = 3.12$, $\nu = 365.8$

Figure 6.2: Illustration of overfitting while optimising $\mathbf{X}$ and (b, ρ, ν). The first three figures show plots of the latent variable against the corresponding event times at successive steps during the optimisation procedure. Initially, in (a), the latent variables strike a balance between satisfying the GPLVM and the WPHM components of the posterior likelihood. However, in (b) it becomes apparent that the latent variables have been adjusted to more tightly fit the WPHM. At the same time, the values of b and ν become larger in absolute magnitude which corresponds to a more sharply peaked event time density (this is reflected in the smaller standard deviations). In (c) the effect becomes even more pronounced. The large parameters in the WPHM lead to numerical inaccuracies in computing the mean and standard deviation of the event time density which can be seen towards the left of the figure. In (d) the event time density for $x = 1$ is plotted after the next optimisation step. The event time density begins to resemble a delta peak centred on the reported event time.



(a) Gamma density with $(\kappa_0, \alpha_0) = (3, 1)$ (b) Gamma density with $(\kappa_1, \alpha_1) = (3, 6)$

Figure 6.3: Prior densities for the shape and scale parameters respectively in the GPLVM-WPHM. These are necessary to avoid implausible solutions that require very large (and intuitively unreasonable) values of the parameters. The prior densities have been chosen such that the scale parameter is suitable for time-to-event measurements in the order of years. Experience suggests that these priors do not have a significant impact on the final solution when compared to flat priors, but they do have a stabilising influence in certain cases by preventing implausible parameter values.

large values of $\mathbf{b}$ and ν which ultimately leads to fatal numerical inaccuracies as the WPHM parameters diverge. For example, the $e^{\mathbf{b}\cdot\mathbf{x}}$ terms will not be evaluated correctly if the argument of the exponential is too large. As the parameters are adjusted to fit the model perfectly the likelihood diverges towards minus infinity. Intuitively the problem is clear: we are trying to explain the survival times with a regression model but are giving ourselves the freedom to adjust both the regression coefficients and the input 'covariates'. We end up finding a solution that fits the model assumptions perfectly. An illustration of the problem is presented in Figure 6.2.

Although the solutions described above are mathematically valid they are unacceptable as explanations for the data we observe. It is unreasonable to suggest that Figure 6.2 (d), for instance, describes a failure probability in reality. Our knowledge of what constitutes a 'plausible' explanation is most appropriately incorporated into our model via prior distributions over model parameters. A suitable prior over ν, for example, would render values $\nu \sim \mathcal{O}(10^2)$ inadmissible.

Following the suggestions of Ibrahim et al. (2001, Section 2.2) we will choose Gamma prior

distributions for the scale and shape parameters in the WPHM:

$$p(\nu|\kappa_0,\alpha_0)=\frac{\nu^{\kappa_0-1}e^{-\nu/\alpha_0}}{\alpha_0^{\kappa_0}\Gamma(\kappa_0)}\quad\text{and}\quad p(\rho|\kappa_1,\alpha_1)=\frac{\rho^{\kappa_1-1}e^{-\rho/\alpha_1}}{\alpha_1^{\kappa_1}\Gamma(\kappa_1)}. \tag{6.14}$$

These are plotted in Figure 6.3. For the regression coefficients we will choose a Gaussian prior:

$$p(\mathbf{b}|\sigma_0)=\frac{e^{-\frac{1}{2\sigma_0^2}\mathbf{b}^2}}{\sqrt{2\pi}\sigma_0}. \tag{6.15}$$

In practice, throughout the simulation studies we will choose $(\kappa_0,\alpha_0)=(3,1)$, $(\kappa_1,\alpha_1)=(3,6)$ and $\sigma_0=2$. These values lead to weakly informative priors which do not significantly alter the inferred parameter values in general, but are sufficient to prevent the nonsensical solutions discussed above. The prior terms that appear in the negative log likelihood (6.6) are (ignoring constant terms)

$$\begin{aligned}-\frac{1}{N}\log p(\mathbf{b})-\frac{1}{N}\log p(\rho)-\frac{1}{N}\log p(\nu)\propto\frac{1}{2N\sigma_0^2}\mathbf{b}^2-\frac{\kappa_0-1}{N}\log\nu+\frac{\nu}{N\alpha_0}\\ -\frac{\kappa_1-1}{N}\log\rho+\frac{\rho}{N\alpha_1}.\end{aligned} \tag{6.16}$$

6.2.4 Implementation

In this section we give practical details of the numerical implementation. As in the case of the GPLVM when we acquire data in the form of covariates and survival outcomes we wish to achieve a number of tasks:

1. Inference of latent variables $\mathbf{X}$ and WPHM parameters $\mathbf{b}$, ρ, and ν.
2. Inference of the hyperparameters.
3. Both of the above in the case of multiple datasets $\{\mathbf{Y}_s\}$.
4. Determining the most appropriate value of q and kernel function.

We will now explain how to perform each task separately.

1. Extraction of latent variables X and WPHM parameters b, ρ, and ν.

We assume for the moment that q, the kernel function, and the hyperparameters are fixed.

1. Initialise $\mathbf{X}$ randomly from a spherical Gaussian density with covariance matrix $\mathbf{I}$, and initialise $\mathbf{b} = \mathbf{0}$, $\nu = 3$ and $\rho = 10$.

2. Minimise the negative log likelihood (6.6) with respect to $\mathbf{X}$ while holding the WPHM parameters fixed. The first time this is done $\mathbf{b} = \mathbf{0}$ so $\mathbf{X}$ will only be influenced by the observed covariates.

3. Minimise (6.6) with respect to the Weibull parameters while $\mathbf{X}$ is fixed to the optimal value from step 2. This is equivalent to simply fitting a WPHM to $\mathbf{X}$ and the survival outcomes.

4. Alternate between Step 2 and Step 3 until a stable solution has been converged upon. In practice we find that ten alterations is sufficient.

2. Inference of the hyperparameters.

The hyperparameters are determined by minimising the negative log hyperparameter likelihood (6.13). As in the case of the GPLVM each evaluation of (6.13) requires relocating the optimal $\mathbf{X}$ and WPHM parameters since they are required for the Laplace approximation.

3. Multiple datasets.

As in the case of the GPLVM each dataset is analysed separately to determine the optimal hyperparameters. These values are then used to train a model that combines multiple datasets by optimising (6.6) with respect to the latent variables and WPHM parameters only.

4. Choice of q and the kernel function

Similarly to the procedure for the GPLVM we can evaluate the minimum negative log hyperparameter likelihood for different values of q in order to determine which is most probable. Within the Bayesian formalism we should integrate over the hyperparameter posterior and compare the posterior over models but that is not feasible in this case. Comparison of the hyperparameter posterior likelihoods is therefore a further approximation.

6.3 Results

In this section we present results from a variety of simulation studies and also some experimental data. We are interested in examining the model's ability to detect and retrieve low dimensional structure under different conditions. We compare the combined GPLVM-WPHM to the unsupervised GPLVM to see if it is advantageous to include survival outcomes when reducing the dimension. We also examine the performance of the model when we combine multiple data sources. We study the effects of overfitting by generating data of various dimensions, and using the GPLVM-WPHM to make predictions for individuals in a validation set. We compute the mean square error (MSE) between the predicted and reported event times, and then study how the MSE changes for data of different dimensions.

Each simulated dataset is randomly generated and consequently when an experiment is repeated the results can be slightly different. To smoothen these statistical fluctuations we repeat the simulated data experiments fifty times and average the results. Finally, we illustrate the practical use of the model by applying it to gene expression signature data with $N = 148$ breast cancer patients from the Guy's Hospital METABRIC dataset (Curtis et al., 2012).

6.3.1 Accuracy of the combined GPLVM-WPHM in comparison to the GPLVM

In this section we want to see if including survival data improves the model's ability to accurately extract the correct low dimensional structure from simulated high dimensional data. To do this we generated fifty datasets from the two dimensional pattern in Figure 5.2 (a). A linear kernel was used with $\beta^2 = 0.01$ and $d = 10$. Survival times were generated from the 'true' latent variables as described in Section 3.5.1. Approximately 20% of the individuals were right censored at random.

For each of the fifty datasets the GPLVM-WPHM was used to extract latent variables with $q = 2$ and the misalignment errors were computed using (5.17), (5.18), and (5.19). The GPLVM was also used to generate a $q = 2$ representation of the high dimensional covariates and misalignment errors were also computed for these representations. We then compared the misalignment errors between the GPLVM-WPHM and the GPLVM. Averaged over the fifty datasets a decrease was observed in the misalignment errors as shown in Table 6.1. We conclude that inclusion of survival data is advantageous since the survival outcomes also contain some information about the latent variables.

β^2	$\mathcal{E}_{radial}$	$\mathcal{E}_{angular}$	$\mathcal{E}_{linear}$
0.1	-7.3%	-5.3%	-6.1%
0.5	-14.5%	-17.1%	-19.5%
1.0	-16.0%	-24.8%	-15.7%

Table 6.1: The average percentage change in misalignment error when the combined GPLVM-WPHM is used instead of the GPLVM. Both models are used to extract a two dimensional latent variable representation of high dimensional simulated data. A decrease in the misalignment error is observed indicating that inclusion of the time-to-event data is beneficial since these data contain useful information about the latent variables. The benefit becomes more apparent as the observed data become noisier.

6.3.2 Integration of multiple sources

Above we saw that including survival outcomes increases the accuracy of the retrieved latent variables. Now we investigate whether including multiple datasets simultaneously leads to similar improvement. We generated one dataset with $d_1 = 10$ and $\beta_1^2 = 0.1$ and a second with $d_2 = 100$ and $\beta_2^2 = 1.0$ (a linear kernel was used in both cases). We computed the misalignment errors after analysing each dataset separately with the GPLVM-WPHM and compare this to the errors obtained after including both datasets simultaneously in the model. The results are shown in Table 6.2 and show that it is beneficial to include both data sources together. These results were averaged over 50 repetitions.

	$\mathcal{E}_{radial}$	$\mathcal{E}_{angular}$	$\mathcal{E}_{linear}$
$\mathbf{Y}_1$ ($d = 10$, $\beta_1^2 = 0.1$)	0.0071	0.0093	0.0270
$\mathbf{Y}_2$ ($d = 100$, $\beta_2^2 = 1.0$)	0.0244	0.0148	0.0509
$\mathbf{Y}_2$ & $\mathbf{Y}_2$	0.0046	0.0052	0.0146

Table 6.2: On the top row are misalignment errors from a model trained on dataset $\mathbf{Y}_1$ alone. Similarly, dataset $\mathbf{Y}_2$ was used to train a model in the middle row. Finally, on the third row are errors from a model that combined both of these datasets simultaneously. Combination of both datasets reduces the error. This is because the overlapping structure in both datasets is reinforced and the signal to noise ratio is increased.

6.3.3 Illustration of overfitting with high dimensional data

In this section want to see the effect of overfitting due to high dimensionality. We generated datasets of different dimensions, each with $N = 200$ individuals from a randomly generated

matrix $\mathbf{X} \in \mathbb{R}^{100\times 2}$ with $q = 2$. Each dataset was split into a training and validation set of equal size. In the high dimensional space we trained a WPHM model on the training individuals and then used the trained model to predict the event time for those individuals in the validation set. We then computed the mean square error (MSE) between the predicted and reported event times (censored individuals were excluded from the validation set).

We then ran the GPLVM-WPHM on the same training data and used the trained model to firstly infer $\mathbf{x}^*$ from $\mathbf{y}^*$ for each individual in the validation set, and subsequently to predict an event time. Again, the MSE is computed and we can compare the MSE in the latent variable space to that obtained in the observed data space. In Table 6.3 we can see that the MSE increases as the dimension of the observed data increases. The results are averaged over fifty datasets.

$d = 10$	$d = 25$	$d = 50$	$d = 100$
+1.2%	+14.7%	+26.6%	+43.4%

Table 6.3: Percentage change in the MSE obtained in the high dimensional covariate spaces compared to the two dimensional latent variable spaces that correspond to each dataset. The MSE is calculated by squaring the difference between predicted and reported event times in validation datasets. We can see that as the dimension of the data increases our predictive accuracy becomes worse; a typical sign of overfitting.

Note that these data were generated with a linear kernel so the increase in MSE is not due to non-linearities induced during the generation of the simulated data. Also, the noise level is relatively low ($\beta^2 = 0.01$) so the observed data are only slightly corrupted with noise. We conclude that the increase in MSE is due to high dimension alone (rather than noise or non-linearities).

We also examined the effect that the noise level has (for fixed d). We can see from Table 6.4 that, in general, the MSE increases with the noise level. The unusually large value for $\beta^2 = 0.5$ is due to an 'outlier' (that is, one particularly bad prediction in the high dimensional space).

6.3.4 Non-linear dimensionality reduction

In this section we investigate the effects that a non-linear mapping between the high and low dimensional spaces can induce. We used the squared exponential kernel to project latent variables with $q = 1$ to $d = 2$. Although the the observed data are not 'high' dimensional they nevertheless lie on a non-linear one dimensional manifold. The dataset was split into

$\beta^2 = 0.01$	$\beta^2 = 0.1$	$\beta^2 = 0.5$	$\beta^2 = 1.0$
+1.2%	+2.7%	+38.0%	+5.67%

Table 6.4: Percentage change in the MSE when computed in the high dimensional space compared to the two dimensional space as a function of the noise level. We can see that as the noise increases we tend to perform worse which is expected. For $\beta^2 = 0.5$ there is an unusually large error which is due to a particularly poor prediction in one of the simulated datasets. Averaging over a greater number of repetitions would help to smoothen these fluctuations but at a greater computational cost.

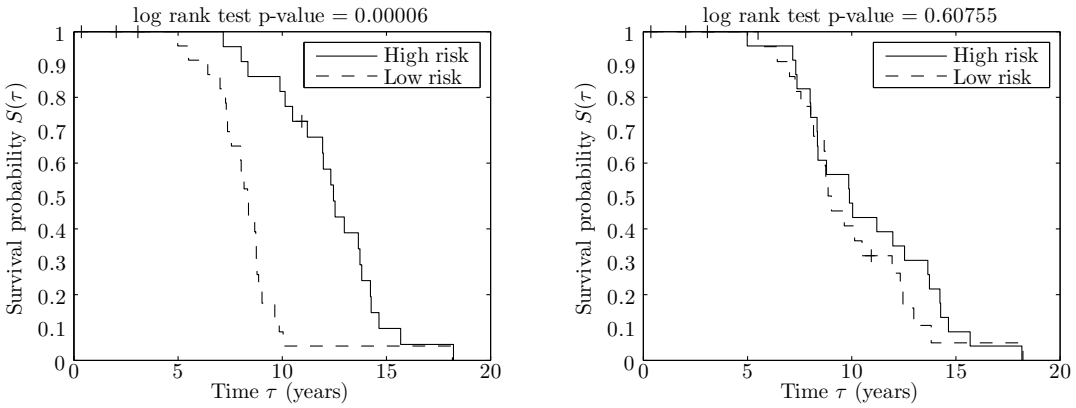
a training set with 48 individuals and a validation set with 48 individuals. We trained a WPHM in the two dimensional training set and then used the GPLVM-WPHM to extract a one dimensional latent variable representation from the same training set. We then used both the trained WPHM and GPLVM-WPHM models to rank the validation individuals according to the values of $\mathbf{b} \cdot \mathbf{x}_i$. The regression coefficients $\mathbf{b}$ come from the trained models and $\mathbf{x}_i$ belong to the 48 validation individuals. The quantity $\mathbf{b} \cdot \mathbf{x}_i$ can be interpreted as a *risk factor* where large positive values indicate high risk and negative values indicate low risk.

In Figure 6.4 we plot Kaplan-Meier curves for the upper and lower quartiles of the ranked validation individuals obtained from the WPHM in two dimensions and the GPLVM-WPHM in one dimension. It is apparent that there is some structure to the low dimensional data. In fact the difference between the survival curves for high and low risk quantiles is statistically significant with a p-value of 0.00006 from a log-rank test. In contrast, there is little difference between the high and low risk groups in the two dimensional space as the structure has been lost due to non-linearities that were induced by the mapping from the original one dimensional space (we found a p-value of 0.60755 using a log-rank test). This illustrates that the GPLVM-WPHM is useful not only for cases where $d > N$ but also cases where non-linear structure can be extracted that may potentially reveal additional patterns of survival.

We can also compare the inferred hyperparameters to those that were used to generate the data which in this case were $(\beta^2, \sigma, l, b, \rho, \nu) = (0.001, 1.00, 1.00, -1.00, 10.0, 10.0)$. The inferred values were $(\beta^2, \sigma, l, b, \rho, \nu) = (0.0006, 1.23, 1.11, -0.68, 9.70, 10.3)$.

6.3.5 Dimensionality detection

Here we give results that illustrate the ability of the GPLVM-WPHM to correctly detect any intrinsic low dimensional structure. This was done by firstly generating low dimensional data

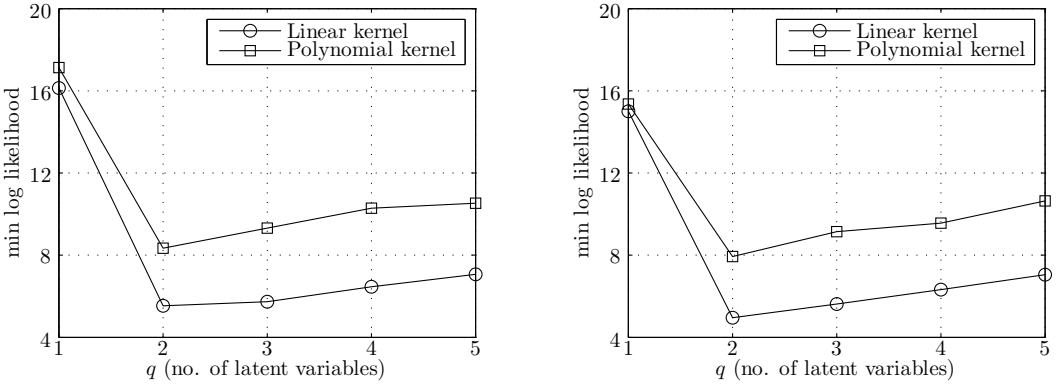


(a) GPLVM-WPHM ($q = 1$) (b) WPHM ($d = 2$)

Figure 6.4: Kaplan-Meier survival curves obtained in the latent variable space $q = 1$ (left) and observed data space with $d = 2$ (right). The data have been generated using the squared exponential kernel and therefore the two dimensional data lie on a one dimensional non-linear manifold. Individuals were split into 'high' and 'low' risk groups by on the basis of risk factors $\mathbf{b} \cdot \mathbf{x}_i$ and Kaplan-Meier curves are plotted for the upper and lower quartiles. We can see a statistically significant difference between the high and low risk groups in the latent variable space but this structure is lost in the two dimensional space.

$\mathbf{X}$ and projecting these into a higher dimensional space. We then trained GPLVM-WPHM models with different values of q and compared the minimum value of the negative log marginal likelihood (6.13). Shown in Figure 6.5 (a) is an example of the model correctly determining that $q = 2$ is the optimal number of latent variables. Additionally, we can compare this to an alternative kernel and we see that the linear kernel (correctly) offers the best description of these data.

In 6.5 (b) we repeat the same experiment using the GPLVM and we see similar results. In fact the GPLVM has a slightly sharper minimum at $q = 2$. One possible explanation for this is that the GPLVM-WPHM is overfitting slightly by using the third latent variable to explain some of the survival outcomes (the three regression coefficients are $b_1 = -3.76$, $b_2 = 0.54$ and $b_3 = 1.19$ which supports this hypothesis).



(a) GPLVM-WPHM (b) GPLVM

Figure 6.5: Dimensionality detection with the GPLVM-WPHM and the GPLVM. Simulated data with $d = 10$, $\beta = 0.1$ and $N = 96$ were generated using a linear kernel along with survival times. The minimum negative log likelihood is plotted as a function of q for both models. Both models correctly detect that two latent variables are optimal. Two different kernel functions are also compared. We infer (correctly) that the linear kernel is the best choice. However the model likelihood ratio between $q = 2$ and $q = 3$ using the GPLVM is 1.94 (that is, two latent variables is almost twice as probable as three), compared to a ratio of 1.23 using the GPLVM-WPHM. A possible explanation for why three latent variables is more probable with the GPLVM-WPHM is that the model is overfitting slightly and using the third latent variable to achieve a better explanation of the survival outcomes (and thus a more probable model).

6.3.6 Experimental data

Finally, we applied the GPLVM-WPHM to a dataset of gene signature scores from breast cancer patients in the Guy's METABRIC dataset[1]. This gene expression dataset from Curtis et al. (2012) was filtered for array intensity, quantile normalised, and batch-corrected for BeadChip. Gene signature scores were calculated as the mean expression of standardised gene expression profiles using previously reported gene lists (Ignatiadis et al., 2012; Palmer et al., 2006; Shipitsin et al., 2007; Patsialou et al., 2012). In total there were $N = 148$ patients without missing data with a total of $d = 26$ signature scores per individual. The cohort was randomly separated into a training and validation set, each containing 74 patients.

The training set was used to train both the GPLVM-WPHM and the WPHM. Multiple instances of the GPLVM-WPHM were trained for each possible value of q and the minimum

[1]We would like to thank Arnie Purushotham, Tony Ng, Anita Grigoriadis and in particular Katherine Lawler for their assistance in accessing and preparing the Guy's METABRIC gene signature score data.

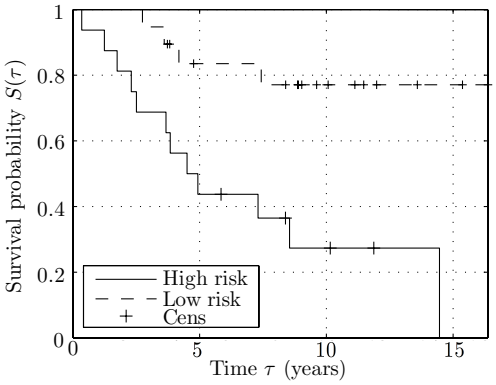
negative log likelihood obtained for each value of q is plotted in Figure 6.6 (c). The optimal number of latent variables is $q = 5$ which indicates substantial redundancy between these gene signatures.

The trained models are then used to classify patients in the validation group into high and low risk quartiles. This is done by ranking all validation patients according to the values of $\mathbf{b} \cdot \mathbf{x}_i$. We then produced Kaplan-Meier survival curves for the upper and lower quartiles of the ranked validation patients. In Figure 6.6 (b) we show the curves obtained using the WPHM with $d = 26$. In Figure 6.6 (a) we show curves obtained from the GPLVM-WPHM with $q = 2$. There is clearly a greater separation when we use the reduced dimension representation of these data. Using a log-rank test we found that there was a statistically significant difference between the K-M curves for high and low risk groups (with a p-value of 0.00208). The same p-value in the high dimensional space was found to be 0.12465. We found that $q = 2$ gave the best separation, which suggests that although $q = 5$ is the optimal solution either not all of those five latent variables are relevant to the validation group or overfitting begins to occur when $q > 2$.

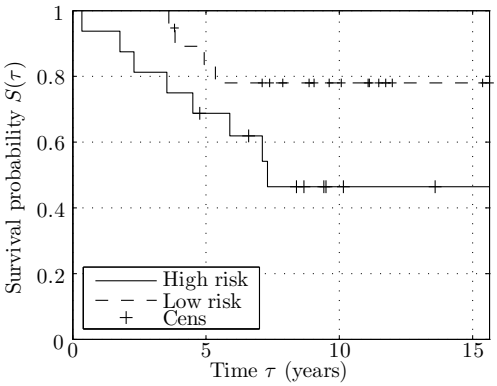
6.4 Discussion

The proposed GPLVM-WPHM offers a novel method of reducing the dimension of survival data in a manner that simultaneously includes the high dimensional covariates and survival outcomes. An interesting question is whether this approach is advantageous to performing an unsupervised dimensionality reduction and survival analysis separately. Our results from simulation studies illustrate that including survival data is worthwhile and leads to more accurate retrieval of low dimensional structure. We also showed that combining information from multiple datasets gives better performance than analysing each dataset on its own. We conclude that in general it is desirable to include all of the available information at the same time since overlapping structure will be reinforced and more robustly detected.

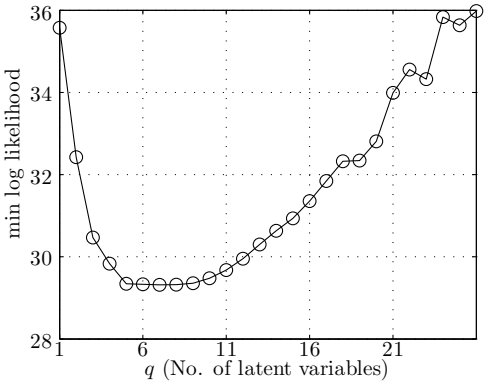
The primary motivation behind developing a dimensionality reduction method was to ultimately increase the accuracy with which we can make predictions for new individuals. The intuition behind this is that by reducing the ratio of covariates to individuals we are more likely to detect genuine patterns in the data. Our results show that reducing the dimension can indeed lead to a significant improvement in predictive accuracy as the effects of overfitting are diminished. In addition, our model can be applied to data that may not be high dimensional but in which the data lie on a non-linear lower dimensional manifold. We used our model



(a) Kaplan-Meier plot with $q = 2$



(b) Kaplan-Meier plot with $d = 26$



(c) Dimensionality detection

Figure 6.6: Results from our analysis of the METABRIC experimental gene expression signature score data. Kaplan-Meier curves for upper and lower quartiles (ranked according to the values of $\mathbf{b} \cdot \mathbf{x}_i$ using the GPLVM-WPHM) in the low dimensional space are shown in (a). Similar curves produced in the original $d = 26$ dimensional space (using the WPHM) are shown in (b). Both of these curves correspond to the validation set of patients. Our ability to predict low and high risk individuals is substantially improved when carrying out regression in the low dimensional latent variable space. The p-values obtained using a log-rank test are 0.00208 and 0.12465 for the GPLVM-WPHM and the WPHM respectively. Figure (c) shows the negative log likelihood for different values of q (the number of latent variables) and illustrates an intrinsic 5 dimensional structure to these data.

to extract non-linear low dimensional structure from simulated data and recover structure to the data that would otherwise have been lost.

Application to real gene expression data illustrate that the GPLVM-WPHM can be of

practical benefit. By comparing the likelihood of models with different numbers of latent variables we were able to infer that there was significant redundancy between the covariates and that by eliminating this redundancy we could achieve a statistically significant difference between high and low risk patient groups that was not observed in the original data.

Future work could involve combining the GPLVM with more sophisticated survival analysis models, such as the GP regression methods developed in the first half of this thesis. See also Lu and Li (2008); Martino et al. (2011); Vanhatalo et al. (2013) for examples of models that allow for flexible non-linear covariate effects in the hazard rate. It would be straightforward to combine alternative models with the GPLVM provided we can obtain the second order partial derivatives which are required for the Laplace approximation of the marginal likelihood. Another research direction would be to apply some of the sparse GP regression techniques in order to reduce the computational burden and speed up the numerical implementation.

Chapter 7

Discussion and Conclusion

In this thesis we have attempted to address two contemporary challenges that biomedical data pose: that of potentially non-linear relationships between covariates and survival outcomes and that of high dimensional datasets. We approached the first problem by applying Gaussian process (GP) regression to survival data. GP regression is sufficiently flexible to capture non-linear covariate effects on survival outcomes. Our approach avoids any assumptions on what form the hazard rates take and thereby imposes fewer structural assumptions on what form the data must take. Using Bayesian methodology we construct a full likelihood function that can easily incorporate any combination of censored or truncated observations. Hyperparameter optimisation is done using the Laplace approximation. Results from both simulated and real data indicate that the model is capable of inferring non-linear relationships between the covariates and time-to-event variables. When we applied our GP regression method to experimental gene expression data we found non-linear patterns and consequently our model vastly out-performed the more traditional Weibull proportional hazards model (WPHM).

A natural extension of this work was to the case of competing risks by building on previous work on multiple output GP regression. The time-to-event variable for each risk is regarded as one of the multiple outputs and the multiple output GP prior is capable of modelling dependencies between these outputs. The main difference between a typical regression setting with multiple outputs and a competing risks setting is that for each individual we observe at most one output, and we know that all of the other outputs must be greater than this reported output. Despite these differences we found that multiple output GP regression performs well on simulated competing risks data. Our approach allows the model to assume whether the risks are dependent or not. Results indicate that assuming dependence between risks can be

advantageous as knowledge of one risk can be used to make more accurate predictions of what the other risks are doing. This approach also benefits from the advantages outlined above, namely the ability to infer non-linear relationships and to incorporate censored and truncated observations easily (although this was not done in this work). One of the key assumptions that we made is that the joint event time density is conditionally independent given the latent function values. Consequently the joint event time density can be written as a product of univariate Gaussian densities which makes the computation of marginal survival probabilities particularly straightforward. It also means that in the hypothetical scenario where we consider what happens after one or more risks are disabled these marginal survival probabilities have a valid interpretation.

In the second part of this thesis we applied the Gaussian process latent variable model (GPLVM) to high dimensional survival data. High dimensional data suffer from the phenomenon of overfitting, where statistical models tend to pick up spurious patterns that fail to generalise to unseen data. This makes it challenging to establish associations between covariates and survival outcomes and to make accurate predictions for new patients. Our goal was to use the GPLVM to generate a low dimensional representation of the observed data and thereby diminish the effects of overfitting. In addition the model provides a way to combine several datasets by expressing them in terms of the same low dimensional latent variables. We first constructed the Laplace approximation of the marginal likelihood. This required the elimination of certain symmetries that existed in the latent variable space. An added bonus of this is that the latent variable representation is always unique. The Laplace approximation of the marginal likelihood was used to optimise model hyperparameters and to compare models with different numbers of latent variables, thereby allowing us to determine the intrinsic dimensionality of a dataset.

Survival outcomes were included by combining the GPLVM with the WPHM. A key assumption of independence between the high dimensional covariates and the survival outcomes conditional on the latent variables was made. This meant it was relatively straightforward in practice to combine the two models. We conducted extensive simulation studies to see what effect the dimension of a dataset has on our ability to make predictions for new individuals. We found that overfitting does indeed lead to a degradation in predictive performance but that the combined GPLVM-WPHM offers a viable route that avoids some of these effects. We also produced evidence that it is beneficial to combine survival analysis with dimensionality reduction simultaneously rather than performing each step separately. Similarly, we showed that combining multiple datasets simultaneously leads to better performance than analysing each

on their own. The GPLVM-WPHM can also be used to extract non-linear low dimensional structure from data that may potentially offer additional insights into survival probabilities. Furthermore, results from the analysis of real gene expression data revealed additional structure to the data in the latent variable space and illustrate the practical usefulness of our model.

In this thesis we hope to have illustrated the usefulness of flexible inference methods such as GP regression which would typically be associated with the 'machine learning' community to the field of survival analysis. By combining the relatively recent work on GP regression and the GPLVM with more firmly established survival analysis methods we hope to have made a useful contribution to the available repertoire of statistical tools that can be applied to survival data. We believe that these methods can provide viable routes to tackle the increasingly complex and high dimensional biomedical data streams currently being produced.

Bibliography

Aalen, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

Abbring, J. H. and Van den Berg, G. J. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):701–710, 2003.

Alvarez, M. A. and Lawrence, N. D. Computationally efficient convolved multiple output Gaussian processes. *The Journal of Machine Learning Research*, 12:1459–1500, 2011.

Andersen, P. K. and Perme, M. P. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99, 2010.

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, 41(3):861–870, 2012.

Bair, E. and Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4), 2004.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473), 2006.

Bernoulli, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour le prévenair. *Historie avec le Mémoires, Académie Royal des Sciences*, 1760.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21(2):263–275, 2002.

Beyersmann, J., Allignol, A., and Schumacher, M. *Competing Risks and Multistate Models with R*. Springer, 2012.

Boyle, P. and Frean, M. Dependent Gaussian processes. *Advances in neural information processing systems*, 17:217–224, 2005.

Chen, K., Jin, Z., and Ying, Z. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002.

Cheng, S. C., Wei, L. J., and Ying, Z. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

Clayton, D. and Cuzick, J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)*, pages 82–117, 1985.

Coolen, A. C. C. and Holmberg, L. *Principles of Survival Analysis (manuscript in preparation)*. Oxford University Press, 2014.

Cox, D. R. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

Crowder, M. *Multivariate Survival Analysis and Competing Risks*. Chapman & Hall/CRC, 2012.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Caldas, C., Aparicio, S., Brenton, J. D., Ellis, I., Huntsman, D., Pinder, S., Purushotham, A., Murphy, L., Bardwell, H., Ding, Z., Jones, L., Liu, B., Papatheodorou, I., Sammut, S. J., Wishart, G., Chia, S., Gelmon, K., Speers, C., Watson, P., Blamey, R., Green, A., Macmillan, D., Rakha, E., Gillett, C., Grigoriadis, A., di Rinaldis, E., Tutt, A., Parisien, M., Troup, S., Chan, D., Fielding, C., Maia, A.-T., McGuire, S., Osborne, M., Sayalero, S. M., Spiteri, I., Hadfield, J., Bell, L., Chow, K., Gale, N., Kovalik, M., Ng, Y., Prentice, L., Tavaré, S., Markowetz, F., Langerø d, A., Provenzano, E., and Bø rresen Dale, A.-L. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, pages 1–7, 2012.

Damianou, A., Ek, C. H., Titsias, M., and Lawrence, N. D. Manifold relevance determination. In *Proceedings of the 29th International Conference in Machine Learning*, 2012.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009.

Di Serio, C. The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study. *Lifetime data analysis*, 3(2):99–122, 1997.

Dickman, P. W., Sloggett, A., Hills, M., and Hakulinen, T. Regression models for relative survival. *Statistics in medicine*, 23(1):51–64, 2004.

Ek, C. H., Rihan, J., Torr, P. H. S., Rogez, G., and Lawrence, N. D. Ambiguity modeling in latent spaces. In *Machine Learning for Multimodal Interaction*, pages 62–73. Springer, 2008.

Ek, C. H., Torr, P. H. S., and Lawrence, N. D. Gaussian process latent variable models for human pose estimation. In *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, volume 4892, pages 132–143. Springer-Verlag Berlin, 2007.

Eleftheriadis, S., Rudovic, O., and Pantic, M. Shared Gaussian process latent variable model for multi-view facial expression recognition. In *Advances in Visual Computing*, pages 527–538. Springer, 2013.

Engler, D. and Li, Y. Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical applications in genetics and molecular biology*, 8(1):1–22, 2009.

Esteve, J., Benhamou, E., Croasdale, M., and Raymond, L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in medicine*, 9(5):529–538, 1990.

Fahrmeir, L. and Kneib, T. *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press, 2011.

Fine, J. P., Ying, Z., and Wei, L. G. On the linear transformation model for censored data. *Biometrika*, 85(4):980–986, 1998.

Fine, J. P. Regression modeling of competing crude failure probabilities. *Biostatistics*, 2(1): 85–97, 2001.

Fine, J. P. and Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

Finkelstein, D. M. A proportional hazards model for interval-censored failure time data. *Biometrics*, pages 845–854, 1986.

Gal, Y., van der Wilk, M., and Rasmussen, C. E. Distributed variational inference in sparse Gaussian process regression and latent variable models. *arXiv preprint arXiv:1402.1389*, 2014.

Gao, X., Wang, X., Tao, D., and Li, X. Supervised Gaussian process latent variable model for dimensionality reduction. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 41:425–34, 2011.

Goeman, J. J. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, 2010.

Goetghebeur, E. and Ryan, L. Semiparametric regression analysis of interval-censored data. *Biometrics*, 56(4):1139–1144, 2000.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, pages 1498–1507, 1998.

Hakulinen, T. and Tenkanen, L. Regression analysis of relative survival rates. *Appl Stat*, 36 (3):309–317, 1987.

Heckman, J. J. and Honoré, B. E. The identifiability of the competing risks model. *Biometrika*, 76(2):325–330, 1989.

Higdon, D. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, 2002.

Hougaard, P. *Analysis of Multivariate Survival Data*. Springer-Verlag New York, Inc., 2000.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. *Bayesian Survival Analysis*. Springer Science+Buisness New York, 2001.

Ignatiadis, M., Singhal, S. K., Desmedt, C., Haibe-Kains, B., Criscitiello, C., Andre, F., Loi, S., Piccart, M., Michiels, S., and Sotiriou, C. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: A pooled analysis. *Journal of Clinical Oncology*, pages 1–10, 2012.

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. Random survival forests for high-dimensional data. *Statistical analysis and data mining*, 4(1):115–132, 2011.

Joensuu, H., Vehtari, A., Riihimäki, J., Nishida, T., Steigen, S. E., Brabec, P., Plank, L., Nilsson, B., Cirilli, C., Braconi, C., Bordoni, A., Magnusson, M. K., Linke, Z., Sufliarsky, J., Federico, M., Jonasson, J. G., Tos, A. P. D., and Rutkowski, P. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology*, 13(3):265–274, 2012.

Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2002.

Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Klein, J. P. and Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data, Second Edition*. Springer Science+Buisness Media, LLC, 2003.

Komárek, A. and Lesaffre, E. The regression analysis of correlated interval-censored data illustration using accelerated failure time models with flexible distributional assumptions. *Statistical Modelling*, 9(4):299–319, 2009.

Kooperberg, C. and Clarkson, D. B. Hazard regression with interval-censored data. *Biometrics*, pages 1485–1494, 1997.

Lambert, P. C., Dickman, P. W., Nelson, C. P., and Royston, P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in medicine*, 29(7-8):885–895, 2010.

Law, C. G. and Brookmeyer, R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine*, 11(12):1569–1578, 1992.

Lawrence, N. D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

Lawrence, N. D. Learning for larger datasets with the Gaussian process latent variable model. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007.

Lindsey, J. K. A study of interval censoring in parametric regression models. *Lifetime Data Analysis*, 4(4):329–354, 1998.

Lu, W. and Li, L. Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*, 9(4):658–667, 2008.

Martino, S., Akerkar, R., and Rue, H. Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38:514 – 528, 2011.

Menzel, D. H. *Fundamental Formulas of Physics*. Dover Publications, Inc. New York, 1960.

Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, pages 951–959, 1992.

Palmer, C., Diehn, M., Alizadeh, A., and Brown, P. O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, 7:115, 2006.

Pan, W. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56(1):199–203, 2000.

Park, M. Y. and Hastie, T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

Patsialou, A., Wang, Y., Lin, J., Whitney, K., Goswami, S., Kenny, P. A., and Condeelis, J. S. Selective gene expression profiling of migratory tumor cells in vivo predicts clinical outcome in breast cancer patients. *Breast Cancer Research*, 14, 2012.

Perme, M. P., Henderson, R., and Stare, J. An approach to estimation in relative survival regression. *Biostatistics*, 10(1):136–146, 2009.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook (version: November 15, 2012). http://matrixcookbook.com, 2012.

Peto, R. Experimental survival curves for interval-censored data. *Applied Statistics*, pages 86–91, 1973.

Quiñonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rabinowitz, D., Tsiatis, A., and Aragon, J. Regression with interval-censored data. *Biometrika*, 82(3):501–513, 1995.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Rencher, A. *Methods of Multivariate Analysis*. Wiley-Interscience, 2002.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Giltnane, J. M. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

Royston, P. and Lambert, P. C. *Flexible Parametric Survival Analysis using Stata: Beyond the Cox Model*. Stata Press, 2011.

Satten, G. A. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83(2):355–370, 1996.

Savitsky, T., Vannucci, M., and Sha, N. Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Statistical Science*, 26(1):130–149, 2011.

Shipitsin, M., Campbell, L. L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M., Halushka, M. K., Sukumar, S., Parker, L. M., Anderson, K. S., Harris, L. N., Garber, J. E., Richardson, A. L., Schnitt, S. J., Nikolsky, Y., Gelman, R. S., and Polyak, K. Molecular definition of breast tumor heterogeneity. *Cancer Cell*, 11(3):259–73, 2007.

Shon, A. P., Grochow, K., Hertzmann, A., and Rao, R. P. N. Learning shared latent structure for image synthesis and robotic imitation. *Advances in Neural Information Processing Systems*, 18:1233, 2006.

Sinha, D., Chen, M.-H., and Ghosh, S. K. Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, 55(2):585–590, 1999.

Snelson, E. and Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.

Snelson, E., Rasmussen, C. E., and Ghahramani, Z. Warped gaussian processes. *Advances in neural information processing systems*, 16:337–344, 2004.

Sohn, I., Kim, J., Jung, S.-H., and Park, C. Gradient lasso for Cox proportional hazards model. *Bioinformatics*, 25(14):1775–1781, 2009.

Sparling, Y. H., Younes, N., Lachin, J. M., and Bautista, O. M. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, 7(4):599–614, 2006.

Tibshirani, R. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

Titsias, M. and Lawrence, N. D. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, 2010.

Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Turnbull, B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

Urtasun, R. and Darrell, T. Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 927–934, New York, NY, USA, 2007. ACM.

Vaida, F. and Xu, R. Proportional hazards model with random effects. *Statistics in medicine*, 19(24):3309–3324, 2000.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(1):1175–1179, 2013.

Vapnik, V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

Vaupel, J. W., Manton, K. G., and Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

Verweij, P. J. M. and Van Houwelingen, H. C. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436, 1994.

Wienke, A. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC biostatistics series, 2011.

Witten, D. M. and Tibshirani, R. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51, 2010.

Zhang, M. and Davidian, M. "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, 64(2):567–576, 2008.

Zhang, Y., Hua, L., and Huang, J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37(2):338–354, 2010.

Zheng, M. and Klein, J. P. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

Appendix A

Partial derivatives for GP regression on survival data

In this appendix we derive the first and second order partial derivatives of the likelihood functions corresponding to the GP regression methods developed in Chapters 3 and 4. The partial derivatives are required for gradient based optimisation of the likelihood and construction of the Laplace approximation of the marginal likelihood. For clarity we will refer to the relevant section of each chapter and rewrite the corresponding negative log likelihood from each section.

A.1 GP regression with a single risk

Here we derive the first and second order partial derivatives required in Section 3.3.3. The negative log likelihood function is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log p(t_i|f_i) - \frac{1}{N}\sum_{i:\Delta_i=0}\log S(t_i|f_i) + \frac{1}{2N}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}) + \frac{1}{2N}\log|\mathbf{K}|. \tag{A.1}$$

First order partial derivatives are

$$\frac{\partial}{\partial f_i}\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{k:\Delta_k=1}\frac{\partial}{\partial f_i}\log p(t_k|f_k) - \frac{1}{N}\sum_{k:\Delta_k=0}\frac{\partial}{\partial f_i}\log S(t_k|f_k) + \frac{1}{N}(\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}))_i \tag{A.2}$$

where

$$\frac{\partial}{\partial f_i}\log p(t_k|f_k) = \delta_{\Delta_i,1}\delta_{ik}\beta^{-2}(t_k - f_k) \tag{A.3}$$

and

$$\begin{aligned}\frac{\partial}{\partial f_i}\log S(t_k|f_k) &= \delta_{\Delta_i,0}\delta_{ik}\frac{1}{S(t_k|f_k)}\int_{t_k}^{\infty}\mathrm{d}s\,\frac{\partial}{\partial f_i}\frac{e^{-\frac{1}{2\beta^2}(s-f_k)^2}}{\sqrt{2\pi}\beta}\\ &= \delta_{\Delta_i,0}\delta_{ik}\frac{1}{S(t_k|f_k)}\frac{1}{\sqrt{2\pi}\beta^3}\int_{t_k}^{\infty}\mathrm{d}s\,(s-f_k)e^{-\frac{1}{2\beta^2}(s-f_k)^2}\\ &= \delta_{\Delta_i,0}\delta_{ik}\frac{1}{S(t_k|f_k)}\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{1}{2\beta^2}(t_k-f_k)^2}.\end{aligned} \tag{A.4}$$

Second order partial derivatives are

$$\begin{aligned}\frac{\partial^2}{\partial f_i\partial f_j}\mathcal{L}(\mathbf{f}) &= -\frac{1}{N}\delta_{ij}\sum_{k:\Delta_k=0}\frac{\partial^2}{\partial f_i^2}\log p(t_k|f_k) - \frac{1}{N}\delta_{ij}\sum_{k:\Delta_k=1}\frac{\partial^2}{\partial f_i^2}\log S(t_k|f_k) + \frac{1}{N}\mathbf{K}_{ij}^{-1}\\ &= \frac{1}{N}(\mathbf{W}+\mathbf{K}^{-1})_{ij}\end{aligned} \tag{A.5}$$

where the diagonal matrix $\mathbf{W}$ is defined by $\mathbf{W}_{ii} = -\frac{\partial^2}{\partial f_i^2}\log p(D|\mathbf{f})$ with

$$\frac{\partial^2}{\partial f_i^2}\log p(t_k|f_k) = -\delta_{\Delta_i,1}\delta_{ik}\beta^{-2} \tag{A.6}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial f_i^2}\log S(t_k|f_k) = \delta_{\Delta_i,0}\delta_{ik}\Bigg[&-\left(\frac{1}{S(t_k|f_k)}\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{1}{2\beta^2}(t_k-f_k)^2}\right)^2\\ &+\frac{(t_k-f_k)}{\beta^2}\left(\frac{1}{S(t_k|f_k)}\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{1}{2\beta^2}(t_k-f_k)^2}\right)\Bigg].\end{aligned} \tag{A.7}$$

Note that the elements of $\mathbf{W}$ are non negative because $S(t|f)$ is log concave. This implies that $\mathbf{H}$ is positive definite which we expect at the minimum of $\mathcal{L}(\mathbf{f})$. To see that $S(t|f)$ is log concave we note that

$$S(t|f) = \int_t^{\infty}\mathrm{d}s\,\frac{e^{-\frac{1}{2\beta^2}(s-f)^2}}{\sqrt{2\pi}\beta} = \int_{t-f}^{\infty}\mathrm{d}s\,\frac{e^{-\frac{1}{2\beta^2}s^2}}{\sqrt{2\pi}\beta} = \int_{-\infty}^{f-t}\mathrm{d}s\,\frac{e^{-\frac{1}{2\beta^2}s^2}}{\sqrt{2\pi}\beta}$$

which is the cumulative distribution function for a Gaussian which is log concave.

A.2 GP regression with interval censored data

The first and second order partial derivatives required in Section 3.3.6 are calculated here. The negative log likelihood function is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log[S(t_i^l|f_i) - S(t_i^u|f_i)] - \frac{1}{N}\sum_{i:\Delta_i=0}\log S(t_i|f_i) - \frac{1}{N}\log p(\mathbf{f}|\mathbf{X}). \tag{A.8}$$

For compactness we define the interval $I_k = (t^l, t^u)$ and

$$\psi(I_k|f_k) = S(\tau_i^l|f_k) - S(\tau_i^u|f_k) = \int_{t^l}^{t^u} \mathrm{d}s\, \frac{e^{-\frac{1}{2\beta^2}(s-f_k)^2}}{\sqrt{2\pi}\beta}. \tag{A.9}$$

The first order partial derivatives are

$$\begin{aligned}
\frac{\partial}{\partial f_i}\log\psi(I_k|f_k) &= \frac{1}{\psi(I_k|f_k)}\frac{\partial}{\partial f_i}\int_{t^l}^{t^u}\mathrm{d}s\,\frac{e^{-\frac{1}{2\beta^2}(s-f_k)^2}}{\sqrt{2\pi}\beta} \\
&= \frac{1}{\psi(I_k|f_k)}\frac{\partial}{\partial f_i}\int_{f_k-t^u}^{f_k-t^l}\mathrm{d}s\,\frac{e^{-\frac{1}{2\beta^2}s^2}}{\sqrt{2\pi}\beta} \\
&= \frac{1}{\psi(I_k|f_k)}\left(\frac{e^{-\frac{1}{2\beta^2}(t^l-f_k)^2}}{\sqrt{2\pi}\beta} - \frac{e^{-\frac{1}{2\beta^2}(t^u-f_k)^2}}{\sqrt{2\pi}\beta}\right).
\end{aligned} \tag{A.10}$$

The second order partial derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial f_i^2}\log\psi(I_k|f_k) = &-\left[\frac{1}{\psi(I_k|f_k)}\left(\frac{e^{-\frac{1}{2\beta^2}(t^l-f_k)^2}}{\sqrt{2\pi}\beta} - \frac{e^{-\frac{1}{2\beta^2}(t^u-f_k)^2}}{\sqrt{2\pi}\beta}\right)\right]^2 \\
&+ \frac{1}{\psi(I_k|f_k)}\left(\frac{(t^l-f_k)}{\beta^2}\frac{e^{-\frac{1}{2\beta^2}(t^l-f_k)^2}}{\sqrt{2\pi}\beta} - \frac{(t^u-f_k)}{\beta^2}\frac{e^{-\frac{1}{2\beta^2}(t^u-f_k)^2}}{\sqrt{2\pi}\beta}\right).
\end{aligned} \tag{A.11}$$

The partial derivatives $\partial^2/\partial f_i \partial f_j \log \Psi = 0$ for $i \neq j$. If we define $h^u = (t^u - f)/\beta\sqrt{2}$ and $h^l = (t^l - f)/\beta\sqrt{2}$ then

$$\psi(I_k|f_k) = \frac{1}{2}\text{erfc}(h^l) - \frac{1}{2}\text{erfc}(h^u) \tag{A.12}$$

$$\frac{\partial}{\partial f_i}\log\psi(I_k|f_k) = \frac{2}{\sqrt{2\pi}\beta}\frac{e^{-(h^l)^2} - e^{-(h^u)^2}}{\text{erfc}(h^l) - \text{erfc}(h^u)} \tag{A.13}$$

$$\frac{\partial^2}{\partial f_i^2}\log\psi(I_k|f_k) = -\left(\frac{2}{\sqrt{2\pi}\beta}\frac{e^{-(h^l)^2} - e^{-(h^u)^2}}{\text{erfc}(h^l) - \text{erfc}(h^u)}\right)^2 + \frac{2}{\sqrt{\pi}\beta}\frac{h^l e^{-(h^l)^2} - h^u e^{-(h^u)^2}}{\text{erfc}(h^l) - \text{erfc}(h^u)}, \tag{A.14}$$

and we can use the asymptotic expansion of the complementary error function to avoid any numerical difficulties (see Section 3.3.7).

A.3 The Joensuu GP hazard rate model

Here we calculate the first and second order partial derivatives required in Section 3.4.3. The negative log likelihood is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\Big[\log\lambda_0(\tau_i) + f(\mathbf{x}_i)\Big] + \frac{1}{N}\sum_{i=1}^{N}\Lambda_0(\tau_i)e^{f(\mathbf{x}_i)} + \frac{1}{2N}\mathbf{f}\cdot\mathbf{K}^{-1}\mathbf{f} + \frac{1}{2N}\log|\mathbf{K}| + \frac{1}{2}\log 2\pi. \tag{A.15}$$

First order derivatives are

$$\frac{\partial}{\partial f_i}\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\delta_{1,\Delta_i} + \frac{1}{N}\Lambda_0(\tau_i)e^{f(\mathbf{x}_i)} + \frac{1}{N}(\mathbf{K}^{-1}\mathbf{f})_i. \tag{A.16}$$

Second order partial derivatives are

$$\begin{aligned}\frac{\partial^2}{\partial f_i \partial f_j}\mathcal{L}(\mathbf{f}) &= \frac{1}{N}\delta_{ij}\Lambda_0(\tau_i)e^{f(\mathbf{x}_i)} + \frac{1}{N}\mathbf{K}^{-1}_{ij} \\ &= \frac{1}{N}(\mathbf{W} + \mathbf{K}^{-1})_{ij}\end{aligned} \tag{A.17}$$

where $\mathbf{W}$ is a diagonal matrix defined by $\mathbf{W}_{ii} = \Lambda_0(t_i)e^{f(\mathbf{x}_i)}$.

A.4 The Joensuu GP hazard rate model with interval censoring

Here we calculate the first and second order partial derivatives required in Section 3.4.4. The negative log likelihood function is

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log S(\tau_i|f_i) - \frac{1}{N}\sum_{i:\Delta_i=0}^{N}\log[S(\tau_i^l|f_i) - S(\tau_i^u|f_i)] + \frac{1}{2N}\mathbf{f}\cdot\mathbf{K}^{-1}\mathbf{f} + \frac{1}{2N}\log|\mathbf{K}| + \frac{1}{2}\log 2\pi. \tag{A.18}$$

The first order partial derivatives can be obtained from

$$-\frac{1}{N}\sum_{k:\Delta_k=1}\frac{\partial}{\partial f_i}\big[-\Lambda_0(\tau_k)e^{-f(\mathbf{x}_k)}\big] = \frac{1}{N}\Lambda_0(\tau_i)e^{f(\mathbf{x}_i)} \tag{A.19}$$

and

$$\sum_{k:\Delta_k=0}\frac{\partial}{\partial f_i}\log[S(\tau_k^l) - S(\tau_k^u)] = \frac{1}{S(\tau_i^l) - S(\tau_i^u)}\Big(-\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}e^{-\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}} + \Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}e^{-\Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}}\Big). \tag{A.20}$$

Second order partial derivatives are given by

$$-\frac{1}{N}\sum_{k:\Delta_k=1}\frac{\partial^2}{\partial f_i\partial f_j}\big(-\Lambda_0(\tau_k)e^{-f(\mathbf{x}_k)}\big) = \frac{\delta_{ij}}{N}\Lambda_0(\tau_k)e^{f(\mathbf{x}_i)} \tag{A.21}$$

and

$$\begin{aligned}\sum_{k:\Delta_k=0}\frac{\partial^2}{\partial f_i\partial f_j}\log[S(\tau_i^l) - S(\tau_i^u)] &= -\delta_{ij}\left(\frac{\partial}{\partial f_i}\log[S(\tau_i^l) - S(\tau_i^u)]\right)^2 \\ &+ \frac{\delta_{ij}}{[S(\tau_i^l) - S(\tau_i^u)]}\left(-\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}e^{-\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}} + \left(\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}\right)^2 e^{-\Lambda_0(\tau_i^l)e^{f(\mathbf{x}_i)}}\right. \\ &\left. + \Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}e^{-\Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}} - \left(\Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}\right)^2 e^{-\Lambda_0(\tau_i^u)e^{f(\mathbf{x}_i)}}\right).\end{aligned} \tag{A.22}$$

Note that (A.20) and (A.22) can be problematic numerically and the approximations discussed in Section 3.4.4 are required.

A.5 GP regression with competing risks

Here we derive the first and second order partial derivatives required in Section 4.4.2. The negative log likelihood function (with two risk and independent right censoring) is

$$\begin{aligned}\mathcal{L}(\mathbf{f}) = &-\frac{1}{N}\sum_{i:\Delta_i\neq 1}\log S_i^1(t_i|f_i^1) - \frac{1}{N}\sum_{i:\Delta_i\neq 2}\log S_i^2(t_i|f_i^2) - \frac{1}{N}\sum_{i:\Delta_i=1}\log p_i(t_i|f_i^1)\\ &-\frac{1}{N}\sum_{i:\Delta_i=2}\log p_i(t_i|f_i^2) + \frac{1}{2N}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}) + \log 2\pi + \frac{1}{2N}\log|\mathbf{K}|. \quad \text{(A.23)}\end{aligned}$$

The first order partial derivatives are

$$\begin{aligned}\frac{\partial}{\partial f_i^r}\mathcal{L}(\mathbf{f}) = &-\frac{1}{N}\sum_{k:\Delta_k\neq 1}\frac{\partial}{\partial f_i^r}\log S_k^1(t_k|f_k^1) - \frac{1}{N}\sum_{k:\Delta_k\neq 2}\frac{\partial}{\partial f_i^r}\log S_k^2(t_k|f_k^2)\\ &-\frac{1}{N}\sum_{k:\Delta_k=1}\frac{\partial}{\partial f_i^r}\log p_k(t_k|f_k^1) - \frac{1}{N}\sum_{k:\Delta_k=2}\frac{\partial}{\partial f_i^r}\log p_k(t_k|f_k^2)\\ &+\frac{1}{N}[\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta})]_i \quad \text{(A.24)}\end{aligned}$$

where

$$\frac{\partial}{\partial f_i^r}\log p_k(t_k|f_k^q) = \delta_{ik}\delta_{pq}\beta_q^{-2}(t_k - f_k^q) \quad \text{(A.25)}$$

and

$$\begin{aligned}\frac{\partial}{\partial f_i^r}S_k^q(t_k|f_k^q) &= \delta_{ik}\delta_{pq}\frac{1}{S_k^q(t_k|f_k^q)}\int_{t_k}^{\infty}\mathrm{d}s\,\frac{\partial}{\partial f_i^r}\frac{e^{-\frac{1}{2\beta_q^2}(s-f_k^q)^2}}{\sqrt{2\pi}\beta_q}\\ &= \delta_{ik}\delta_{pq}\frac{1}{S_k^q(t_k|f_k^q)}\frac{1}{\sqrt{2\pi}\beta_q^3}\int_{t_i}^{\infty}\mathrm{d}s\,(s-f_k^q)e^{-\frac{1}{2\beta_q^2}(s-f_k^q)^2}\\ &= \delta_{ik}\delta_{pq}\frac{1}{S_k^q(t_k|f_k^q)}\frac{1}{\sqrt{2\pi}\beta_q}e^{-\frac{1}{2\beta_q^2}(t_k-f_k^q)^2}. \quad \text{(A.26)}\end{aligned}$$

Second order partial derivatives are (where $\partial^2/\partial f_j^r \partial f_i^q = 0$ for $r \neq q$)

$$\frac{\partial^2}{\partial f_j^r \partial f_i^r}\mathcal{L}(\mathbf{f}) = \frac{1}{N}(\mathbf{W} + \mathbf{K}^{-1})_{ij} \tag{A.27}$$

with

$$\begin{aligned}\mathbf{W}_{ij} = &- \sum_{k:\Delta_k \neq 1} \frac{\partial^2}{\partial f_i^r \partial f_j^r} \log S_k^1(t_k|f_k^1) - \sum_{k:\Delta_k \neq 2} \frac{\partial^2}{\partial f_i^r \partial f_j^r} \log S_k^2(t_k|f_k^2) \\ &- \sum_{k:\Delta_k = 1} \frac{\partial^2}{\partial f_i^r \partial f_j^r} \log p_k(t_k|f_k^1) - \sum_{k:\Delta_k = 2} \frac{\partial^2}{\partial f_i^r \partial f_j^r} \log p_k(t_k|f_k^2).\end{aligned} \tag{A.28}$$

The matrix $\mathbf{W}$ is diagonal since

$$\begin{aligned}\frac{\partial^2}{\partial f_i^r \partial f_j^r} \log S_k^q(t_k|f_k^q) = \delta_{ik}\delta_{jk} &\left(\frac{1}{S_k^q(t_k|f_k^q)} \frac{1}{(2\pi\beta_q^2)^{1/2}} e^{-\frac{1}{2\beta_q^2}(t_k - f_k^q)^2} \right)^2 \\ &- \frac{(t_k - f_k^q)}{\beta_q^2} \left(\frac{1}{S_k^q(t_k|f_k^q)} \frac{1}{(2\pi\beta_q^2)^{1/2}} e^{-\frac{1}{2\beta_q^2}(t_k - f_k^q)^2} \right)\end{aligned} \tag{A.29}$$

and

$$\frac{\partial^2}{\partial f_i^r \partial f_j^r} \log p_k^q(t_k|f_k^q) = \delta_{ik}\delta_{jk}\beta_q^{-2}. \tag{A.30}$$

Appendix B

Partial derivatives of the GPLVM and the GPLVM-WPHM

In this appendix we derive the first and second order partial derivatives of the likelihood functions corresponding to the GPLVM and the combined GPLVM-WPHM developed in Chapters 5 and 6. The partial derivatives are required for gradient based optimisation algorithms and construction of the Laplace approximation. Again, for clarity we will refer to the relevant section of each chapter and rewrite the corresponding negative log likelihood from each section.

B.1 The GPLVM

In Section 5.2.2 we require the first and second order partial derivatives of the negative log likelihood:

$$\mathcal{L}(\mathbf{X}) = \sum_{s=1}^{S} \left(\frac{d_s}{2N} \mathrm{tr}(\mathbf{K}_s^{-1} \mathbf{S}_s) + \frac{d_s}{2N} \log |\mathbf{K}_s| + \frac{d_s}{2} \log 2\pi \right) - \frac{1}{N} \log p(\mathbf{X}). \tag{B.1}$$

The following identities are used (Petersen and Pedersen, 2012):

$$\frac{\partial |\mathbf{K}|}{\partial \mathbf{K}} = |\mathbf{K}|\mathbf{K}^{-1} \tag{B.2}$$

$$\frac{\partial \mathrm{tr}(\mathbf{A}\mathbf{K}^{-1}\mathbf{B})}{\partial \mathbf{K}} = -(\mathbf{K}^{-1}\mathbf{B}\mathbf{A}\mathbf{K}^{-1})^{\mathrm{T}} \tag{B.3}$$

$$\frac{\mathbf{K}^{-1}_{kl}}{\partial \mathbf{K}_{ij}} = \mathbf{K}^{-1}_{ki}\mathbf{K}^{-1}_{jl}. \tag{B.4}$$

The following is true for any type of kernel function

$$\frac{\partial}{\partial \mathbf{X}}\mathcal{L}(\mathbf{X}) = \sum_{s=1}^{S}\sum_{i,j=1}^{N}\frac{\partial \mathcal{L}}{\partial \mathbf{K}^s_{ij}}\frac{\partial \mathbf{K}^s_{ij}}{\partial \mathbf{X}}, \tag{B.5}$$

where from (B.2, B.3)

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}_s} = -\frac{d_s}{2N}\mathbf{K}_s^{-1}\mathbf{S}_s\mathbf{K}_s^{-1} + \frac{d_s}{2N}\mathbf{K}_s^{-1}. \tag{B.6}$$

In what follows we drop the index s for clarity and derive the partial derivatives for the linear, squared exponential and polynomial kernels.

The linear kernel

The linear kernel is defined by

$$\mathbf{K}_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j + \beta^2\delta_{ij}. \tag{B.7}$$

First order partial derivatives are

$$\frac{\partial}{\partial \mathbf{X}}\mathcal{L}(\mathbf{X}) = -\frac{d}{N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{X} + \frac{d}{N}\mathbf{K}^{-1}\mathbf{X}. \tag{B.8}$$

The following expressions are needed to construct the second order partial derivatives

$$\begin{aligned}\frac{\partial}{\partial x_{p\nu}}(-\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{X})_{r\mu} = &-(\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1})_{rp}\delta_{\mu\nu} + (\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{X})_{p\mu}(\mathbf{K}^{-1}\mathbf{X})_{r\nu} + \\ &+ (\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{X})_{r\nu}(\mathbf{K}^{-1}\mathbf{X})_{p\mu} + (\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1})_{rp}(\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X})_{\nu\mu} + \\ &+ (\mathbf{X}^T\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{X})_{\nu\mu}(\mathbf{K}^{-1})_{rp}\end{aligned} \tag{B.9}$$

and

$$\frac{\partial}{\partial x_{p\nu}}(\mathbf{K}^{-1})_{r\mu} = (\mathbf{K}^{-1})_{rp}\delta_{\mu\nu} - (\mathbf{K}^{-1}\mathbf{X})_{r\nu}(\mathbf{K}^{-1}\mathbf{X})_{p\mu} - (\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X})_{\nu\mu}(\mathbf{K}^{-1})_{rp}. \tag{B.10}$$

The squared exponential kernel

The squared exponential kernel function is defined by

$$\mathbf{K}_{ij} = \sigma e^{-\frac{l}{2}(\mathbf{x}_i - \mathbf{x}_j)^2} + \beta^2\delta_{ij}. \tag{B.11}$$

Evaluation of (B.5) takes $\mathcal{O}(N^2)$ operations to compute. However, for i, j and r distinct

$$\frac{\partial \mathbf{K}_{ij}}{\partial x_{r\mu}} = 0 \tag{B.12}$$

$$\frac{\partial \mathbf{K}_{ir}}{\partial x_{r\mu}} = \frac{\partial \mathbf{K}_{ri}}{\partial x_{r\mu}} = l\sigma(x_{i\mu} - x_{r\mu})e^{-\frac{l}{2}(\mathbf{x}_i - \mathbf{x}_r)^2} = l(x_{i\mu} - x_{r\mu})\mathbf{K}_{ir}. \tag{B.13}$$

This allows us to compute (B.5) with $\mathcal{O}(N)$ operations since from (B.5) we can write

$$\frac{\partial \mathcal{L}}{\partial x_{r\mu}} = 2\sum_i \frac{\partial \mathcal{L}}{\partial \mathbf{K}_{ir}}\frac{\partial \mathbf{K}_{ir}}{\partial x_{r\mu}}. \tag{B.14}$$

Second order partial derivatives are obtained by differentiating (B.14) again to obtain

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_{p\nu}\partial x_{r\mu}} = 2\sum_i^N \Bigg\{ &\frac{\partial}{\partial x_{p\nu}}\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1}\right]_{ir}[l(x_{i\mu} - x_{r\mu})\mathbf{K}_{ir}] \\ &+ \left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1}\right]_{ir}\frac{\partial}{\partial x_{p\nu}}[l(x_{i\mu} - x_{r\mu})\mathbf{K}_{ir}]\Bigg\}.\end{aligned} \tag{B.15}$$

On the first line there will be two terms. Beginning with the first term and using (B.3) and (B.12, B.13) and we can write

$$\begin{aligned}\frac{\partial}{\partial x_{p\nu}}(\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1})_{ir} &= \frac{\partial}{\partial x_{p\mu}}\sum_{tl}\mathbf{K}^{-1}_{it}\mathbf{S}_{tl}\mathbf{K}^{-1}_{lr} \\ &= \sum_{tl}\mathbf{K}^{-1}_{it}\mathbf{S}_{tl}\left[\frac{\partial}{\partial x_{p\nu}}\mathbf{K}^{-1}_{lr}\right] + \sum_{tl}\left[\frac{\partial}{\partial x_{p\nu}}\mathbf{K}^{-1}_{it}\right]\mathbf{S}_{tl}\mathbf{K}^{-1}_{lr}\end{aligned} \tag{B.16}$$

which can be simplified to

$$\frac{\partial}{\partial x_{p\mu}}(\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1})_{ir} = \sum_{k=1}^{N} \frac{\partial \mathbf{K}_{pk}}{\partial x_{p\nu}} \left(-[\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}]_{ik}\mathbf{K}^{-1}_{pr} - [\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}]_{ip}\mathbf{K}^{-1}_{kr} \right.$$
$$\left. -[\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}]_{pr}\mathbf{K}^{-1}_{ik} - [\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}]_{kr}\mathbf{K}^{-1}_{ip} \right). \quad \text{(B.17)}$$

From (B.4) and (B.12, B.13) we can write the second term as

$$\frac{\partial}{\partial x_{p\nu}}\mathbf{K}^{-1}_{ir} = \sum_{k=1}^{N} \frac{\partial \mathbf{K}_{pk}}{\partial x_{p\nu}} \left(-\mathbf{K}^{-1}_{ik}\mathbf{K}^{-1}_{pr} - \mathbf{K}^{-1}_{ip}\mathbf{K}^{-1}_{kr} \right). \quad \text{(B.18)}$$

Finally, the term on the second line of (B.15) is

$$\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1} \right]_{pr} \left[l^2(x_{p\mu} - x_{r\mu})(x_{r\nu} - x_{p\nu}) + l\delta_{\mu\nu} \right] \mathbf{K}_{pr} \qquad \text{when } i = p \text{ and } r \neq p$$
$$\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1} \right]_{ip} \left[l^2(x_{i\mu} - x_{p\mu})(x_{i\nu} - x_{p\nu}) - l\delta_{\mu\nu} \right] \mathbf{K}_{ip} \qquad \text{when } i \neq p \text{ and } r = p$$

and zero otherwise.

The polynomial kernel

The polynomial kernel of degree α is defined by

$$\begin{aligned} \mathbf{K}_{ij} &= \sigma(1 + \mathbf{x}_i \cdot \mathbf{x}_j)^{\alpha} + \beta^2\delta_{ij} \\ &= \sigma \sum_{n=0}^{\alpha} \binom{\alpha}{n} (\mathbf{x}_i \cdot \mathbf{x}_j)^n + \beta^2\delta_{ij}. \end{aligned} \quad \text{(B.19)}$$

The partial derivatives of $\mathbf{K}$ with respect to $\mathbf{X}$ are

$$\frac{\partial \mathbf{K}_{ij}}{\partial x_{r\mu}} = 0 \quad \text{(B.20)}$$

$$\frac{\partial \mathbf{K}_{ir}}{\partial x_{r\mu}} = \frac{\partial \mathbf{K}_{ri}}{\partial x_{r\mu}} = \sigma \sum_{n=1}^{\alpha} \binom{\alpha}{n} n(\mathbf{x}_i \cdot \mathbf{x}_r)^{n-1} x_{i\mu} \quad \text{(B.21)}$$

$$\frac{\partial \mathbf{K}_{rr}}{\partial x_{r\mu}} = 2\sigma \sum_{n=1}^{\alpha} \binom{\alpha}{n} n(\mathbf{x}_r \cdot \mathbf{x}_r)^{n-1} x_{r\mu}. \quad \text{(B.22)}$$

Insertion into (B.5) yields

$$\frac{\partial \mathcal{L}}{\partial x_{r\mu}} = 2 \sum_i \frac{\partial \mathcal{L}}{\partial \mathbf{K}_{ir}} \frac{\partial \mathbf{K}_{ir}}{\partial x_{r\mu}}. \tag{B.23}$$

Second order partial derivatives

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_{p\nu} \partial x_{r\mu}} = 2 \sum_i^N \Bigg\{ & \frac{\partial}{\partial x_{p\nu}} \left[-\frac{d}{2N} \mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1} + \frac{d}{2N} \mathbf{K}^{-1} \right]_{ir} \left[\sigma \sum_{n=1}^{\alpha} \binom{\alpha}{n} n (\mathbf{x}_i \cdot \mathbf{x}_r)^{n-1} x_{i\mu} \right] \\ & + \left[-\frac{d}{2N} \mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1} + \frac{d}{2N} \mathbf{K}^{-1} \right]_{ir} \frac{\partial}{\partial x_{p\nu}} \left[\sigma \sum_{n=1}^{\alpha} \binom{\alpha}{n} n (\mathbf{x}_i \cdot \mathbf{x}_r)^{n-1} x_{i\mu} \right] \Bigg\}.\end{aligned} \tag{B.24}$$

As in the case of the squared exponential kernel the first line contains two terms. They are

$$\begin{aligned}\frac{\partial}{\partial x_{p\nu}} (\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1})_{ir} &= \frac{\partial}{\partial x_{p\mu}} \sum_{tl} \mathbf{K}^{-1}_{it} \mathbf{S}_{tl} \mathbf{K}^{-1}_{lr} \\ &= \sum_{tl} \mathbf{K}^{-1}_{it} \mathbf{S}_{tl} \left[\frac{\partial}{\partial x_{p\nu}} \mathbf{K}^{-1}_{lr} \right] + \sum_{tl} \left[\frac{\partial}{\partial x_{p\nu}} \mathbf{K}^{-1}_{it} \right] \mathbf{S}_{tl} \mathbf{K}^{-1}_{lr}\end{aligned} \tag{B.25}$$

which can be simplified (with $\alpha = 2$) to

$$\begin{aligned}\frac{\partial}{\partial x_{p\mu}} (\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1})_{ir} = & \sum_{k=1}^{N} 2\sigma x_{k\nu} (1 + \mathbf{x}_k \mathbf{x}_p)) \Big(-[\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1}]_{ik} \mathbf{K}^{-1}_{pr} - [\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1}]_{ip} \mathbf{K}^{-1}_{kr} \\ & - [\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1}]_{pr} \mathbf{K}^{-1}_{ik} - [\mathbf{K}^{-1} \mathbf{S} \mathbf{K}^{-1}]_{kr} \mathbf{K}^{-1}_{ip} \Big).\end{aligned} \tag{B.26}$$

The second term is

$$\frac{\partial}{\partial x_{p\nu}} \mathbf{K}^{-1}_{ir} = \sum_{k=1}^{N} \Big[- \mathbf{K}^{-1}_{ik} \mathbf{K}^{-1}_{pr} - \mathbf{K}^{-1}_{ip} \mathbf{K}^{-1}_{kr} \Big] (2\sigma x_{k\nu} (1 + \mathbf{x}_k \cdot \mathbf{x}_p)). \tag{B.27}$$

Terms from the second line of (B.24) are (with $\alpha = 2$)

$$\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1}\right]_{ip} 2\sigma x_{i\mu}x_{i\nu} \qquad \text{when } i \neq p \text{ and } r = p$$

$$\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1}\right]_{pp} 2\sigma\left(\delta_{\mu\nu}(1 + \mathbf{x}_p^2) + 2x_{p\mu}x_{p\nu}\right) \qquad \text{when } i = p \text{ and } r = p$$

$$\left[-\frac{d}{2N}\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1} + \frac{d}{2N}\mathbf{K}^{-1}\right]_{pr} 2\sigma\left(\delta_{\mu\nu}(1 + \mathbf{x}_p \cdot \mathbf{x}_r) + x_{p\mu}x_{r\nu}\right) \qquad \text{when } i = p \text{ and } r \neq p$$

and zero otherwise.

Implementational details

The sums over k in (B.17, B.18) and (B.26, B.27) and the sums over i in (B.15, B.24) can be eliminated by performing vectorised operations over appropriately defined matrices in Matlab. Matrices such as $\mathbf{X}\mathbf{X}^T$ and $\mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}$ can be computed outside any loops. Since the Hessian matrix is symmetric it is necessary only to compute $Nq(Nq-1)/2$ partial derivatives.

B.2 The GPLVM predictive distribution

In Section 5.2.4 we require the partial derivatives of

$$\mathcal{L}(\mathbf{x}^*) = \frac{1}{N}\sum_{s=1}^{S}\left(\frac{1}{2\kappa_s^2}(\mathbf{y}_s^* - \mathbf{m}_s)^2 + \frac{d_s}{2}\log 2\pi + d_s \log \kappa_s\right). \tag{B.28}$$

The first order partial derivatives are

$$\begin{aligned}\frac{\partial}{\partial x_\mu^*}\mathcal{L}(\mathbf{x}^*) = -\frac{1}{N\kappa^3}\left(\frac{\partial \kappa}{\partial x_\mu^*}\right)\sum_{\nu=1}^{d}(y_\nu^* - m_\nu)^2 \\ - \frac{1}{N\kappa^2}\sum_{\nu=1}^{2}(y_\nu^* - m_\nu)\left(\frac{\partial m_\nu}{\partial x_\mu^*}\right) + \frac{d}{N\kappa}\left(\frac{\partial \kappa}{\partial x_\mu^*}\right).\end{aligned} \tag{B.29}$$

For the linear kernel

$$\frac{\partial m_\nu}{\partial x_\mu^*} = \mathbf{x}_\mu \cdot \mathbf{K}^{-1}\mathbf{y}_\nu \tag{B.30}$$

and

$$\frac{\partial \kappa}{\partial x_\mu^*} = 2x_\mu^* - \mathbf{x}_\mu \cdot \mathbf{K}^{-1}\mathbf{k} - \mathbf{k} \cdot \mathbf{K}^{-1}\mathbf{x}_\mu \tag{B.31}$$

where $\mathbf{x}_\mu \in \mathbb{R}^{N\times 1}$ is the μth column of $\mathbf{X}$. For the squared exponential kernel these are

$$\frac{\partial m_\nu}{\partial x_\mu^*} = -l \sum_{i=1}^{N} (x_\mu^* - x_{i\mu}) k_i [\mathbf{K}^{-1}\mathbf{y}_\nu]_i \tag{B.32}$$

and

$$\frac{\partial \kappa}{\partial x_\mu^*} = 2l \sum_{i=1}^{N} (x_\mu^* - x_{i\mu}) k_i [\mathbf{K}^{-1}\mathbf{k}]_i, \tag{B.33}$$

where $k_i = [\mathbf{k}(\mathbf{x}^*, \mathbf{X})]_i$. For the polynomial kernel we have

$$\frac{\partial m_\nu}{\partial x_\mu^*} = 2 \sum_{i=1}^{N} (1 + \mathbf{x}^* \cdot \mathbf{x}_i) x_{i\mu} [\mathbf{K}^{-1}\mathbf{y}_\nu]_i \tag{B.34}$$

and

$$\frac{\partial \kappa}{\partial x_\mu^*} = 4(1 + \mathbf{x}^* \cdot \mathbf{x}^*) x_\mu^* - 4 \sum_{i=1}^{N} (1 + \mathbf{x}^* \cdot \mathbf{x}_i) x_{i\mu} [\mathbf{K}^{-1}\mathbf{k}]_i. \tag{B.35}$$

B.3 The combined GPLVM-WPHM model

In Section 6.2.2 we require first and second order partial derivatives of the negative log likelihood:

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mathbf{b}, \rho, \nu) = \sum_{s=1}^{S} \left(\frac{d_s}{2N} \mathrm{tr}(\mathbf{K}_s^{-1}\mathbf{S}_s) + \frac{d_s}{2N} \log |\mathbf{K}_s| + \frac{d_s}{2} \log 2\pi \right) \\ - \frac{1}{N} \sum_{i:\Delta_i=1} [\log \lambda_0(\tau_i) + \mathbf{b} \cdot \mathbf{x}_i] + \frac{1}{N} \sum_{i=1}^{N} \Lambda_0(\tau_i) e^{\mathbf{b}\cdot\mathbf{x}_i} \\ - \frac{1}{N} \log p(\mathbf{b}) - \frac{1}{N} \log p(\rho) - \frac{1}{N} \log p(\nu).\end{aligned} \tag{B.36}$$

The first line contains terms from the GPLVM and these partial derivatives are given above in Section B.1. The third line consists of prior terms belonging to the WPHM. Partial derivatives for these terms are derived in Appendix C. What remains to calculate are partial derivatives

of WPHM terms on the second line with respect to $\mathbf{X}$. We define

$$\mathcal{L}_D(\mathbf{b}, \rho, \nu) = -\frac{1}{N} \sum_{i:\Delta_i=1} [\log \lambda_0(\tau_i) + \mathbf{b} \cdot \mathbf{x}_i] + \frac{1}{N} \sum_{i=1}^{N} \Lambda_0(\tau_i) e^{\mathbf{b}\cdot\mathbf{x}_i}. \tag{B.37}$$

Then

$$\frac{\partial \mathcal{L}_D}{\partial x_{r\mu}} = \frac{1}{N} \left[-\delta_{1,\Delta_r} b_\mu + b_\mu \Lambda_0(\tau_r) e^{\mathbf{b}\cdot\mathbf{x}_r} \right] \tag{B.38}$$

and

$$\frac{\partial^2 \mathcal{L}_D}{\partial x_{p\eta} \partial x_{r\mu}} = \frac{1}{N} \delta_{pr} b_\mu b_\eta \Lambda_0(\tau_r) e^{\mathbf{b}\cdot\mathbf{x}_r}. \tag{B.39}$$

Using (C.2)

$$\frac{\partial \mathcal{L}_D}{\partial x_{r\mu} \partial b_\eta} = \frac{1}{N} \left\{ \Lambda_0(\tau_r) x_{r\eta} b_\mu e^{\mathbf{b}\cdot\mathbf{x}_r} + \delta_{\mu\eta} \left[\Lambda_0(\tau_r) e^{\mathbf{b}\cdot\mathbf{x}_r} - \delta_{1,\Delta_r} \right] \right\}. \tag{B.40}$$

Using (C.3) and (C.4) we obtain

$$\frac{\partial \mathcal{L}_D}{\partial x_{r\mu} \partial \rho} = -\frac{1}{N} \frac{\nu}{\rho} \left(\frac{\tau_r}{\rho} \right)^{\nu} b_\mu e^{\mathbf{b}\cdot\mathbf{x}_r} \tag{B.41}$$

$$\frac{\partial \mathcal{L}_D}{\partial x_{r\mu} \partial \nu} = \frac{1}{N} (\log \tau_r - \log \rho) \left(\frac{\tau_r}{\rho} \right)^{\nu} b_\mu e^{\mathbf{b}\cdot\mathbf{x}_r}. \tag{B.42}$$

Due to the parameterisation (C.9) we will use the following partial derivatives in practice:

$$\frac{\partial \mathcal{L}_D}{\partial x_{r\mu} \partial \tilde{\rho}} = \frac{\partial \mathcal{L}_D}{\partial x_{r\mu} \partial \rho} \frac{\partial \rho}{\partial \tilde{\rho}} \quad \text{with} \quad \frac{\partial \rho}{\partial \tilde{\rho}} = \frac{e^{\tilde{\rho}}}{1 + e^{\tilde{\rho}}}. \tag{B.43}$$

Appendix C

Partial derivatives of the Weibull proportional hazards model

In this appendix we give the first and second order partial derivatives of the Weibull proportional hazards model (WPHM). These are needed for gradient based optimisation of the likelihood function and construction of the Laplace approximation. In Section 3.4.2 we require partial derivatives of the negative log likelihood for the WPHM. The negative log likelihood is

$$\mathcal{L}(\boldsymbol{\beta}, \rho, \nu) = -\frac{1}{N} \sum_{i:\Delta_i=1} [\log \lambda_0(\tau_i) + \boldsymbol{\beta} \cdot \mathbf{x}_i] + \frac{1}{N} \sum_{i=1}^{N} \Lambda_0(\tau_i) e^{\boldsymbol{\beta} \cdot \mathbf{x}_i} - \frac{1}{N} \log p(\boldsymbol{\beta}) - \frac{1}{N} \log p(\rho) - \frac{1}{N} \log p(\nu). \tag{C.1}$$

The first order partial derivatives are

$$\frac{\partial}{\partial \beta_s} \mathcal{L}(\boldsymbol{\beta}, \rho, \nu) = -\frac{1}{N} \sum_{i:\Delta_i=1} x_{is} + \frac{1}{N} \sum_{i=1}^{N} \Lambda_0(\tau_i) x_{is} e^{\boldsymbol{\beta} \cdot \mathbf{x}_i} \tag{C.2}$$

and

$$\frac{\partial}{\partial\rho}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{N_1}{N}\frac{\nu}{\rho} + \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\Lambda_0(\tau_i)}{\partial\rho}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \tag{C.3}$$

$$\frac{\partial}{\partial\nu}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = -\frac{N_1}{N}\frac{1}{\nu} - \frac{1}{N}\sum_{i:\Delta_i=1}\log(\tau_i/\rho) + \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\Lambda_0(\tau_i)}{\partial\nu}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}. \tag{C.4}$$

We have used

$$\frac{\partial}{\partial\rho}\log\lambda_0(\tau) = -\frac{\nu}{\rho} \tag{C.5}$$

$$\frac{\partial}{\partial\nu}\log\lambda_0(\tau) = \frac{1}{\nu} + \log(\tau/\rho) \tag{C.6}$$

and

$$\frac{\partial\Lambda_0(\tau)}{\partial\rho} = -\frac{\nu}{\rho}\left(\frac{\tau}{\rho}\right)^{\nu} \tag{C.7}$$

$$\frac{\partial\Lambda_0(\tau)}{\partial\nu} = (\log\tau - \log\rho)\left(\frac{\tau}{\rho}\right)^{\nu}. \tag{C.8}$$

Since we require $\rho > 0$ we write it in the form

$$\rho = \log(1 + \rho_{LB} + \exp(\tilde{\rho})) \tag{C.9}$$

where $\tilde{\rho} \in \mathbb{R}$ and $\rho_{LB} \geq 0$ is a lower bound on ρ that can be set manually. This formulation allows the use of unconstrained optimisation functions to be used. The partial derivatives now become

$$\frac{\partial\mathcal{L}}{\partial\tilde{\rho}} = \frac{\partial\mathcal{L}}{\partial\rho}\frac{\partial\rho}{\partial\tilde{\rho}} \quad \text{with} \quad \frac{\partial\rho}{\partial\tilde{\rho}} = \frac{e^{\tilde{\rho}}}{1+e^{\tilde{\rho}}}. \tag{C.10}$$

We also require $\nu > 0$ and the same formulation is used.

Second order partial derivatives are

$$\frac{\partial^2}{\partial\beta_r\partial\beta_s}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{1}{N}\sum_{i=1}^{N}\Lambda_0(\tau_i)x_{is}x_{ir}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \tag{C.11}$$

and

$$\frac{\partial^2}{\partial\rho^2}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = -\frac{N_1}{N}\frac{\nu}{\rho^2} + \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\nu(\nu+1)}{\rho^2}\left(\frac{\tau_i}{\rho}\right)^{\nu}\right]e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \tag{C.12}$$

$$\frac{\partial^2}{\partial\nu^2}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{N_1}{N}\frac{1}{\nu^2} + \frac{1}{N}\sum_{i=1}^{N}(\log\tau_i - \log\rho)^2\left(\frac{\tau_i}{\rho}\right)^{\nu}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}. \tag{C.13}$$

Finally we require

$$\frac{\partial^2}{\partial\nu\partial\rho}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{\partial^2}{\partial\rho\partial\nu}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{N_1}{N}\frac{1}{\rho} - \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\nu}{\rho}(\log\tau_i - \log\rho)\left(\frac{\tau_i}{\rho}\right)^{\nu} + \frac{1}{\rho}\left(\frac{\tau_i}{\rho}\right)^{\nu}\right] \tag{C.14}$$

$$\frac{\partial^2}{\partial\rho\partial\beta_s}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = -\frac{1}{N}\frac{\nu}{\rho}\sum_{i=1}^{N}x_{is}\left(\frac{\tau_i}{\rho}\right)^{\nu}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \tag{C.15}$$

$$\frac{\partial^2}{\partial\nu\partial\beta_s}\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = \frac{1}{N}\sum_{i=1}^{N}(\log\tau_i - \log\rho)x_{is}\left(\frac{\tau_i}{\rho}\right)^{\nu}e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}. \tag{C.16}$$

Since in practice we write the parameters ρ and ν in the form (C.9) the second order partial derivatives are

$$\frac{\partial^2\mathcal{L}}{\partial\tilde{\rho}^2} = \frac{\partial^2\mathcal{L}}{\partial\rho^2}\left(\frac{\partial\rho}{\partial\tilde{\rho}}\right)^2 + \frac{\partial\mathcal{L}}{\partial\rho}\frac{\partial^2\rho}{\partial\tilde{\rho}^2} \quad \text{with} \quad \frac{\partial^2\rho}{\partial\tilde{\rho}^2} = \frac{e^{\tilde{\rho}}}{(1+e^{\tilde{\rho}})^2} \tag{C.17}$$

$$\frac{\partial^2\mathcal{L}}{\partial\tilde{\rho}\partial\tilde{\nu}} = \frac{\partial^2\mathcal{L}}{\partial\tilde{\nu}\partial\tilde{\rho}} = \frac{\partial^2\mathcal{L}}{\partial\rho\partial\nu}\frac{\partial\rho}{\partial\tilde{\rho}}\frac{\partial\nu}{\partial\tilde{\nu}} \tag{C.18}$$

$$\frac{\partial^2\mathcal{L}}{\partial\tilde{\rho}\partial\beta_s} = \frac{\partial^2\mathcal{L}}{\partial\rho\partial\beta_s}\frac{\partial\rho}{\partial\tilde{\rho}} \text{ and } \frac{\partial^2\mathcal{L}}{\partial\tilde{\nu}\partial\beta_s} = \frac{\partial^2\mathcal{L}}{\partial\nu\partial\beta_s}\frac{\partial\nu}{\partial\tilde{\nu}}. \tag{C.19}$$

Prior terms

When we assume the prior distributions (6.14) and (6.15) from Section 6.2.3 we will need to include some additional terms in the first and second order partial derivatives. These are

$$-\frac{1}{N}\frac{\partial}{\partial b_\mu}\log p(\mathbf{b}) = \frac{1}{N\sigma_0^2}b_\mu \tag{C.20}$$

$$-\frac{1}{N}\frac{\partial}{\partial \nu}\log p(\nu) = -\frac{\kappa_0 - 1}{N\nu} + \frac{1}{N\alpha_0} \tag{C.21}$$

$$-\frac{1}{N}\frac{\partial}{\partial \rho}\log p(\rho) = -\frac{\kappa_1 - 1}{N\rho} + \frac{1}{N\alpha_1} \tag{C.22}$$

and

$$-\frac{1}{N}\frac{\partial^2}{\partial b_\mu \partial b_\nu}\log p(\mathbf{b}) = \delta_{\mu\nu}\frac{1}{N\sigma_0^2} \tag{C.23}$$

$$-\frac{1}{N}\frac{\partial^2}{\partial \nu^2}\log p(\nu) = \frac{\kappa_0 - 1}{N\nu^2} \tag{C.24}$$

$$-\frac{1}{N}\frac{\partial^2}{\partial \rho^2}\log p(\rho) = \frac{\kappa_1 - 1}{N\rho^2}. \tag{C.25}$$

Appendix D

Matrix identities and Gaussian integrals

In Chapter 4 we will require some of the results presented here. In particular, we will require the following identity

$$(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}. \tag{D.1}$$

This can be verified by writing

$$\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{A}[\mathbf{B}^{-1}(\mathbf{A} + \mathbf{B})]^{-1} = [(\mathbf{B}^{-1}\mathbf{A} + \mathbf{I})\mathbf{A}^{-1}]^{-1} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}. \tag{D.2}$$

In what follows we will also need the Woodbury identity which is

$$(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}. \tag{D.3}$$

We will now prove the following identity:

$$\mathbf{B} - \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}. \tag{D.4}$$

This can be shown by first using Woodbury's identity to write

$$\begin{aligned} \mathbf{B} - \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} &= \mathbf{B} - \mathbf{B}[\mathbf{B}^{-1} - \mathbf{B}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}]\mathbf{B} \\ &= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}, \end{aligned} \tag{D.5}$$

from which we can see that identity (D.4) follows immediately from (D.1).

In Section 4.3 we use the following result

$$\int \mathrm{d}\mathbf{z}\, e^{-\frac{1}{2}(\mathbf{z}-\mathbf{a})\cdot\mathbf{A}(\mathbf{z}-\mathbf{a})} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{b})\cdot\mathbf{B}(\mathbf{z}-\mathbf{b})} = \frac{(2\pi)^{d/2}}{\sqrt{|\mathbf{A}+\mathbf{B}|}} e^{-\frac{1}{2}(\mathbf{b}-\mathbf{a})\cdot\mathbf{D}(\mathbf{b}-\mathbf{a})} \tag{D.6}$$

where $\mathbf{D} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$. To prove this we write the argument of the exponential on the left hand side as

$$(\mathbf{z}-\mathbf{a})\cdot\mathbf{A}(\mathbf{z}-\mathbf{a}) + (\mathbf{z}-\mathbf{b})\cdot\mathbf{B}(\mathbf{z}-\mathbf{b}) = (\mathbf{z}-\mathbf{c})\cdot\mathbf{C}(\mathbf{z}-\mathbf{c}) + \mathbf{d}. \tag{D.7}$$

Equating both sides we find (and noting that $\mathbf{a}\cdot\mathbf{A}\mathbf{b} = \mathbf{b}\cdot\mathbf{A}\mathbf{a}$ for symmetric $\mathbf{A}$)

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \tag{D.8}$$

$$\mathbf{c} = \mathbf{C}^{-1}(\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b}) \tag{D.9}$$

$$\mathbf{d} = \mathbf{a}\cdot\mathbf{A}\mathbf{a} + \mathbf{b}\cdot\mathbf{B}\mathbf{b} - \mathbf{c}\cdot\mathbf{C}\mathbf{c}. \tag{D.10}$$

We can now write

$$\begin{aligned}\int \mathrm{d}\mathbf{z}\, e^{-\frac{1}{2}(\mathbf{z}-\mathbf{a})\cdot\mathbf{A}(\mathbf{z}-\mathbf{a})} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{b})\cdot\mathbf{B}(\mathbf{z}-\mathbf{b})} &= e^{\frac{1}{2}\mathbf{d}} \int \mathrm{d}\mathbf{z}\, e^{-\frac{1}{2}(\mathbf{z}-\mathbf{c})\cdot\mathbf{C}(\mathbf{z}-\mathbf{c})} \\ &= \frac{(2\pi)^{d/2}}{\sqrt{|\mathbf{C}|}} e^{-\frac{1}{2}\mathbf{d}}.\end{aligned} \tag{D.11}$$

Next, we write

$$\begin{aligned}\mathbf{d} &= \mathbf{a}\cdot\mathbf{A}\mathbf{a} + \mathbf{b}\cdot\mathbf{B}\mathbf{b} - \mathbf{a}\cdot\mathbf{A}\mathbf{C}^{-1}\mathbf{A}\mathbf{a} - \mathbf{a}\cdot\mathbf{A}\mathbf{C}^{-1}\mathbf{B}\mathbf{b} - \mathbf{b}\cdot\mathbf{B}\mathbf{C}^{-1}\mathbf{A}\mathbf{a} - \mathbf{a}\cdot\mathbf{B}\mathbf{C}^{-1}\mathbf{B}\mathbf{b} \\ &= (\mathbf{b}-\mathbf{e})\cdot\mathbf{E}(\mathbf{b}-\mathbf{e}) + \mathbf{g},\end{aligned} \tag{D.12}$$

from which it follows that

$$\begin{aligned}\mathbf{E} &= \mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} \\ &= \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\end{aligned} \tag{D.13}$$

where we have used (D.4) to obtain the second line. It is easy to verify that $\mathbf{E} = \mathbf{A} - \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$ also. We also find that $\mathbf{e} = \mathbf{E}^{-1}\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{a} = \mathbf{a}$. Finally, $\mathbf{g} = \mathbf{a}\cdot\mathbf{A}\mathbf{a} + \mathbf{b}\cdot\mathbf{B}\mathbf{b} -$

$\mathbf{e} \cdot \mathbf{E}\mathbf{e} = 0$. Substituting these into (D.11) we obtain (D.6) as desired.